On the Design of Channel Shortening Demodulators for Iterative Receivers in Linear Vector Channels

Sha Hu and Fredrik Rusek
Department of Electrical and Information Technology
Lund University, Lund, Sweden
{sha.hu,fredrik.rusek}@eit.lth.se

Abstract

We consider the problem of designing demodulators for linear vector channels with memory that use reduced-size trellis descriptions for the received signal. We assume an overall iterative receiver, and for the parts of the signal not covered by the trellis description, we use interference cancelation based on the soft information provided by the outer decoder. In order to reach a trellis description, a linear filter is applied as front-end to compress the signal structure into a small trellis. This process requires three parameters to be designed: (i) the front-end filter, (ii) the feedback filter through which the interference cancelation is done, and (iii) a target response which specifies the trellis. Demodulators of this form have been studied before under then name *channel shortening* (CS), but the interplay between CS and interference cancelation has not been adequately addressed in the literature. In this paper, we analyze two types of CS demodulators that are based on the Forney and Ungerboeck detection models, respectively. The parameters are jointly optimized based on a generalized mutual information (GMI) function. We also introduce a third type of CS demodulator that is in general suboptimal but has closed form solutions for all parameters. Moreover, signal to noise ratio (SNR) asymptotic properties are analyzed and we show that the third CS demodulator asymptotically converges to the optimal CS demodulator in the sense of maximizing the GMI.

This paper will be presented in part at the IEEE 26th annual international symposium on personal, indoor and mobile radio communications (PIMRC) in Hongkong on August 30 - September 2, 2015.

I. Introduction

For intersymbol interference (ISI) channels, Forney [1] showed that the Viterbi Algorithm (VA) [2] implements maximum likelihood (ML) detection. However, the complexity of the VA is exponential in the memory of the channel which prohibits its use in many cases of interest. As a remedy, Falconer and Magee proposed in 1973 the concept of channel shortening [3]. The concept is to filter the received signal with a channel shortening filter so that the effective channel has much shorter duration than the original channel, and then apply the VA to the shorter effective channel.

CS demodulators have a long and rich history, see [3]–[14]. Traditionally, the CS demodulators have been optimized from a minimum mean square error (MMSE) perspective [3]–[12]. Two exceptions from this are the papers [13] and [14]. In [13], the authors attempt to minimize the error probability of an uncoded system which leads to a new notion of posterior equivalence between the target response and the filtered channel. However, since [13] works with uncoded error probabilities, the analysis in [13] does not adequately address the case of coded systems and Shannon capacity properties.

To the best of our knowledge, the first paper that works with capacity-related cost measures is [14]. In [14] the authors consider the achievable rate, in the form of generalized mutual information (GMI) [15]–[19], that the transceiver system can achieve if a CS demodulator is adopted. However, [14] is limited to ISI channels only, and the design method in [14] of the CS demodulator is in fact not always possible to execute. The limitations of [14] were first dealt with in [18], which extended the CS concept to any linear vector channel and resulted in a closed form optimization procedure.

In this paper we generalize the idea in [18] to iterative receivers. With iterative receivers it is reasonable to expect that better performance can be reached by allowing the parameters of the CS demodulator to change in each iteration. A limitation in [18] is that the CS demodulator does not take the prior information into account, rendering its design static in all iterations. We extend the static CS demodulators described in [18] and aim at constructing a CS demodulator that takes soft information provided by the outer decoder into account so that the parameters of the CS demodulator are designed for a particular level of prior knowledge. This procedure includes an interference cancelation mechanism to deal with the signal part that can not be handled by the trellis search. Preliminary results for CS demodulators in iterative receivers are available in [20], but this paper non-trivially advances the state of the art.

A closely related concept is delayed-decision-feedback-sequence-estimation (DDFSE) first investigated in [21]. However, in DDFSE the interference cancelation is done within a single iteration, and not between the iterations of an iterative receiver.

The paper is organized as follows: The linear vector channel model is described in Section II while the general form of the CS demodulators and the iterative receiver structure are introduced in Section III. In Section IV we analyze three types of CS demodulators for finite length linear vector channels and in Section V we deal with ISI channels as asymptotic versions of the results established in Section IV. The SNR asymptotics of the CS demodulators are discussed in Section VI. Numerical results are provided in Section VII and Section VIII summarizes the paper. For improved readability we have deferred some long proofs and derivations to Appendices A-K.

A. Notation

Throughout the paper, a capital bold letter such as "$\boldsymbol{A}$" represents a matrix, a lower case bold letter "$\boldsymbol{a}$" represents a vector and the capital letter "A" represents a number. "$\boldsymbol{A}\prec 0$" means matrix $\boldsymbol{A}$ is negative definite while $\boldsymbol{A}\succ 0$" means $\boldsymbol{A}$ is positive definite. Matrix $\boldsymbol{I}$ represents the identity matrix and in general the dimension will be omitted; when it cannot be understood from the context, we let $\boldsymbol{I}_K$ represent a $K\times K$ identity matrix. Our superscripts have the following meanings: "$*$" is complex conjugate, "T" is matrix transpose, "H" denotes the conjugate transpose of a matrix, "-1" is matrix inverse, "$-\mathrm{T}$" means both matrix inverse and transpose, and "$-\mathrm{H}$" denotes both inverse and conjugate transpose of a matrix. In addition, "$\propto$" means proportional to, "$\mathbb{E}[\,]$" is the expectation operator, "$\mathrm{Tr}(\,)$" takes the trace of a matrix, "$\mathrm{Re}\{\,\}$" returns the real part of a variable, "$\otimes$" is the Kronecker multiplication operator, $\mathrm{vec}(\boldsymbol{A})$ is a column vector containing the columns of matrix $\boldsymbol{A}$ stacked on top of each other, and "$[A,B]$" is the set of integers $\{k\!:\!A\!\le\! k\!\le\! B\}$.

Furthermore, we say that a matrix $\boldsymbol{A}$ is banded within diagonals $[-\nu_1,\nu_2]$ $(\nu_1,\nu_2\!\ge\!0)$, if the

(k,ℓ)th element $A(k,\ell)$ satisfies[1],

$$A(k,\ell)=0,\ \ell-k>\nu_1 \text{ or } k-\ell>\nu_2.$$

Moreover, we define two matrix operators $[\ ]_\nu$ and $[\ ]_{\backslash\nu}$ such that $\boldsymbol{A}=[\boldsymbol{A}]_\nu+[\boldsymbol{A}]_{\backslash\nu}$, $[\boldsymbol{A}]_\nu$ is banded within diagonals $[-\nu,\nu]$ where $[\boldsymbol{A}]_{\backslash\nu}$ is constrained to zero.

II. System Model

We consider linear vector channels according to

$$\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}+\boldsymbol{n} \tag{1}$$

where $\boldsymbol{y}$ is an $N\times 1$ vector of received signal, $\boldsymbol{x}$ is a $K\times 1$ vector comprising unit energy coded symbols that belong to a constellation $\mathcal{X}$, $\boldsymbol{H}$ is an $N\times K$ matrix representing the communication channel and $\boldsymbol{n}$ is zero-mean complex Gaussian noise vector with covariance matrix $N_0\boldsymbol{I}$. Denote x_k as the kth element of $\boldsymbol{x}$ and $\boldsymbol{h}_k$ as the kth column vector of $\boldsymbol{H}$, (1) can be rewritten as

$$\boldsymbol{y}=\sum_{k=0}^{K-1}\boldsymbol{h}_k x_k+\boldsymbol{n}. \tag{2}$$

In an iterative receiver, the feedback from the outer decoder can be utilized in the demodulator to improve the performance. As the outer decoder provides the demodulator with a posteriori probability (APP) and extrinsic information (in terms of bit log-likelihood ratio (LLR)) [22], [23], side information is present about the symbols $\boldsymbol{x}$ and we represent this by the probability mass function $p_k(s)=\mathrm{P}(x_k=s)$, $0\le k\le K-1$. Note that the side-information does not consider the dependency among the symbols, but are symbolwise marginal probabilities. This reflects the situation encountered in iterative receivers with perfect interleaving. In those cases, the prior probabilities provided from previous iterations are assumed independent, i.e., $\mathrm{P}(\boldsymbol{x}=\boldsymbol{s})=\prod p_k(s)$. Due to the perfect interleaving assumption, the demodulator can compute $\hat{\boldsymbol{x}}=[\ \hat{x}_1\ \cdots\ \hat{x}_K\ ]^{\mathrm{T}}=\mathbb{E}(\boldsymbol{x})$ in a per-entry fashion as

$$\hat{x}_k=\sum_{s\in\mathcal{X}} s p_k(s).$$

[1] Note that ν_1 refers to the number of upper diagonals of $\boldsymbol{A}$ that are non-zero. We have this convention in order to subsequently follow standard notation for Toeplitz matrices [42].

Further, the $K\times K$ diagonal matrix $\boldsymbol{P}=\mathbb{E}[\boldsymbol{x}\hat{\boldsymbol{x}}^{\mathrm{H}}]=\mathbb{E}[\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^{\mathrm{H}}]$ that reflects the quality of the side information can be calculated, and the expectations are computed with respect to the prior distribution $p_k(s)$.

The task of the demodulator is to generate soft information about the symbols in $\boldsymbol{x}$ given the observable $\boldsymbol{y}$ and the side information $\{p_k(s)\}$. The optimal demodulator is the maximum-a-posteriori (MAP) demodulator [24], [25] which evaluates the posterior probabilities $\mathrm{P}(x_k\!=\!s|\boldsymbol{y})$. However, the number of leaves of the search tree corresponding to the MAP demodulator is in general $|\mathcal{X}|^K$ which is prohibitive for most practical applications. The purpose of the CS demodulator is to force the signal model to be an lower triangular matrix with only $\nu+1$ $(0\!\leq\!\nu\!<\!K\!-\!1)$ non-zero diagonals by means of a linear filter[2]. ν is referred to as the memory length of the CS demodulator. Then, a BCJR [26] demodulator can be applied over a trellis with $|\mathcal{X}|^\nu$ states. Moreover, since there is side information present about $\boldsymbol{x}$, the parts of $\boldsymbol{H}$ acting as noise can be partly eliminated by means of interference cancelation through the prior mean $\hat{\boldsymbol{x}}$.

Notice that if we set $\nu=K-1$, the search space of CS demodulator is no longer a trellis but corresponds to the original tree and is therefore equivalent to MAP. On the other hand, the linear MMSE demodulator with parallel interference cancelation (LMMSE-PIC) [27]–[29] is a special case of CS demodulation with $\nu=0$. In the LMMSE-PIC, the BCJR is trivial since different symbols are assumed to be independent after the front-end filtering. Therefore the CS demodulator is a generalized framework that includes both the MAP and LMMSE-PIC demodulators. The CS demodulator can also be viewed as an extension of the LMMSE-PIC to include a trellis search, where the parameters for the front-end filter, the interference cancelation and the trellis search process are jointly optimized.

III. The General Form of the CS Demodulator

We state two lemmas first that will be useful later. Lemma 2 can be verified straightforwardly.

Lemma 1. *Let $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ be two $K\!\times\! K$ matrices, $\boldsymbol{A}_1$ is invertible and banded within diagonals $[-\nu,\nu]$. If $[\boldsymbol{A}_1^{-1}]_\nu\!=\![\boldsymbol{A}_2]_\nu$, then*

$$\mathrm{Tr}\big(\boldsymbol{A}_1\boldsymbol{A}_2\big)=\mathrm{Tr}\big(\boldsymbol{I}\big).$$

[2] For finite length linear vector channels such as multi-input multi-output (MIMO) channel, “filtering” means matrix multiplication.

Proof: Let $\boldsymbol{A}_3 = \boldsymbol{A}_2 - \boldsymbol{A}_1^{-1}$, then $[\boldsymbol{A}_3]_\nu = \boldsymbol{0}$ and $\boldsymbol{A}_3 = [\boldsymbol{A}_3]_{\backslash\nu}$. As $\boldsymbol{A}_1 = [\boldsymbol{A}_1]_\nu$, the elements along the main diagonal of $\boldsymbol{A}_1\boldsymbol{A}_3$ are zero. Therefore $\mathrm{Tr}\big(\boldsymbol{A}_1\boldsymbol{A}_2\big) = \mathrm{Tr}\big(\boldsymbol{A}_1(\boldsymbol{A}_1^{-1} + \boldsymbol{A}_3)\big) = \mathrm{Tr}(\boldsymbol{I})$. ■

Lemma 2. *Let $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ be two $K \times K$ matrices that are banded within diagonals $[-\nu_1, \nu_2]$ and $[-\nu_3, \nu_4]$, respectively. Then the product $\boldsymbol{A}_1\boldsymbol{A}_2$ is banded within diagonals $[\max(-(\nu_1 + \nu_3), 1 - K), \min(\nu_2 + \nu_4, K - 1)]$.*

A. System Model of the CS Demodulator

The CS demodulators that we investigate operate on the basis of the following mismatched function

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{x}}) = \exp\big(2\mathrm{Re}\{\boldsymbol{x}^{\mathrm{H}}(\boldsymbol{V}\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{x}})\} - \boldsymbol{x}^{\mathrm{H}}\boldsymbol{G}\boldsymbol{x}\big) \tag{3}$$

instead of the true conditional probability

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{(\pi N_0)^N} \exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2}{N_0}\right). \tag{4}$$

Note that $\tilde{p}(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{x}})$ may not be a valid probability distribution function, but this is irrelevant for demodulation, see [30]. The matrices $\boldsymbol{V}$, $\boldsymbol{R}$ and $\boldsymbol{G}$ are referred to as the front-end filter, interference cancelation matrix and trellis representation matrix, respectively. Without loss of generality, we have absorbed N_0 into $\boldsymbol{V}$, $\boldsymbol{R}$ and $\boldsymbol{G}$. Note that (3) and (4) are equivalent for demodulation if we set $\boldsymbol{V} = \boldsymbol{H}^{\mathrm{H}}/N_0$, $\boldsymbol{R} = \boldsymbol{0}$ and $\boldsymbol{G} = \boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}/N_0$ in which case the CS demodulator represents the MAP demodulator without interference cancelation.

In this paper we will go through three types of CS demodulators that can be expressed in the form (3), but with different constructions of matrices $\boldsymbol{V}$, $\boldsymbol{R}$ and $\boldsymbol{G}$ that represent different views on the domain in which the CS should be performed.

In order to optimize the matrices $(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G})$, we choose to work with the GMI which is an achievable rate for a receiver that operates on the basis of a mismatched version of the channel law. The GMI in nats/channel equals

$$I_{\mathrm{GMI}} = -\mathbb{E}\left[\log \tilde{p}(\boldsymbol{y}|\hat{\boldsymbol{x}})\right] + \mathbb{E}\left[\log \tilde{p}(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{x}})\right] \tag{5}$$

where $\tilde{p}(\boldsymbol{y}|\hat{\boldsymbol{x}}) = (1/\pi^K)\int \tilde{p}(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{x}}) \exp(-\|\boldsymbol{x}\|^2)\mathrm{d}\boldsymbol{x}$ and the expectation is taken over the true statistics of the channel. Notice that while finite constellations $\mathcal{X}$ are almost always used in practice, they are hard to analyze. In order to obtain a mathematically tractable problem, here we use a zero-mean, unit variance, complex Gaussian constellation for each entry of $\boldsymbol{x}$. With complex Gaussian inputs, the trellis discussed earlier has no proper meaning as the number of states is infinite even for finite ν. However, the complex Gaussian assumption is only made in order to design the receiver parameters $(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G})$.

Theorem 1. *With the system model in (3), the GMI defined in (5) reads,*

$$I_{\mathrm{GMI}}(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G}) = \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) - \mathrm{Tr}(\boldsymbol{G}) + 2\mathrm{Re}\big\{\mathrm{Tr}(\boldsymbol{V}\boldsymbol{H}-\boldsymbol{R}\boldsymbol{P})\big\}$$
$$-\mathrm{Tr}\Big((\boldsymbol{I}+\boldsymbol{G})^{-1}\big(\boldsymbol{V}(N_0\boldsymbol{I}+\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}})\boldsymbol{V}^{\mathrm{H}} - 2\mathrm{Re}\big\{\boldsymbol{V}\boldsymbol{H}\boldsymbol{P}\boldsymbol{R}^{\mathrm{H}}\big\} + \boldsymbol{R}\boldsymbol{P}\boldsymbol{R}^{\mathrm{H}}\big)\Big). \quad (6)$$

Here we make the same assumption as in [18] that $\boldsymbol{I}+\boldsymbol{G}$ is positive definite, otherwise the GMI is not well defined. The proof of Theorem 1 is given in Appendix A.

With any parameters $(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G})$, the GMI can be calculated from (6), although they may not be optimal in the sense of maximizing the GMI. We illustrate Theorem 1 with two examples.

Example 1. *Extended Zero-Forcing filter (EZF). We extend the zero-Forcing filter [31] to only partly invert the channel so that a trellis search is necessary after the EZF front-end filter. In view of the CS demodulator, we can select the parameters in (3) as:*

$$\boldsymbol{V} = (\boldsymbol{I}+\boldsymbol{G})(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{H}}, \ \boldsymbol{R} = \boldsymbol{0},$$

and then optimize (6) over G. In order to satisfy the constraint of having a trellis with $|\mathcal{X}|^{\nu}$ states, we should have $\boldsymbol{G} = [\boldsymbol{G}]_{\nu}$. The optimal $\boldsymbol{G}$, in the sense of maximizing (6), will be shown (Theorem 2) to satisfy,

$$[(\boldsymbol{I}+\boldsymbol{G})^{-1}]_{\nu} = N_0[(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1}]_{\nu}.$$

Utilizing Lemma 1, the GMI in (6) for the optimal $\boldsymbol{G}$ equals

$$I_{\mathrm{GMI}} = \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) + \mathrm{Tr}\big(\boldsymbol{I} - N_0(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1}(\boldsymbol{I}+\boldsymbol{G})\big) = \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big).$$

Example 2. *Truncated Matched filter (TMF). As previously mentioned, the MAP demodulator (4) can be written in the form (3) by setting $\boldsymbol{V} = \boldsymbol{H}^{\mathrm{H}}/N_0$, $\boldsymbol{R} = \boldsymbol{0}$ and $\boldsymbol{G} = \boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}/N_0$. The front-end is in this case a matched filter [32] and the BCJR needs to be implemented over the Ungerboeck model [33]. In order to reach a trellis with $|\mathcal{X}|^{\nu}$ states, we can truncate $\boldsymbol{G}$ to its*

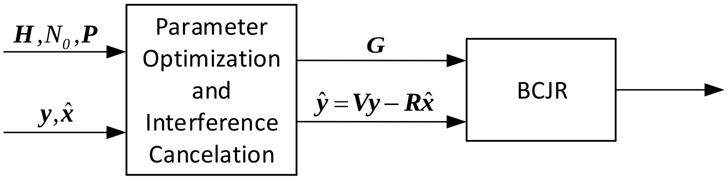
H,N_0,P
y,$\hat{x}$
Parameter Optimization and Interference Cancelation
G
$\hat{y} = Vy - R\hat{x}$
BCJR

Fig. 1: CS demodulator that maximizes the GMI.

center $2\nu+1$ *diagonals, i.e., we can use the following parameters in (3):*

$$\boldsymbol{V} = \boldsymbol{H}^{\mathrm{H}}/N_0, \ \boldsymbol{R} = \boldsymbol{0}, \ \text{and} \ \boldsymbol{G} = [\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}/N_0]_\nu.$$

With these choices, the GMI in (6) equals

$$I_{\mathrm{GMI}} = \log\big(\det(\boldsymbol{I} + [\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}/N_0]_\nu)\big) - \mathrm{Tr}\big(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}\big(N_0\boldsymbol{I} + [\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}]_\nu\big)^{-1}[\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}]_{\setminus\nu}\big).$$

B. Constraints on the Parameter $\boldsymbol{R}$ for the CS Demodulator

Our approach to design a CS demodulator consists of two steps. As illustrated in Figure 1 these are:

- Construction of a signal $\hat{\boldsymbol{y}}$ based on the received signal $\boldsymbol{y}$ and prior mean $\hat{\boldsymbol{x}}$ as $\hat{\boldsymbol{y}} = \boldsymbol{V}\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{x}}$.
- BCJR demodulation of $\hat{\boldsymbol{y}}$ operating on a reduced number of states $|\mathcal{X}|^\nu$.

This procedure is fully analogous to an LMMSE-PIC demodulator which first subtracts the interference, applies a Wiener filter and concludes by a BCJR demodulator that operates with a diagonal matrix $\boldsymbol{G}$.

As mentioned earlier, optimization of the demodulator will be made on the basis of GMI which is evaluated for the statistical model of the tuple $(\boldsymbol{x}, \hat{\boldsymbol{y}})$. The statistical behavior of $(\boldsymbol{x}, \hat{\boldsymbol{y}})$ may be superior to that of the original $(\boldsymbol{x}, \boldsymbol{y})$ as the former tuple corresponds to a statistically different channel than the true one. Hence, the GMI may very well exceed the channel capacity. Moreover, the computed value of GMI may have little relevance for the performance of the transceiver system. In order for GMI to have bearing on performance, it is critical to put constraints on the matrix $\boldsymbol{R}$ as our next example will show.

Example 3. *Let the system model be*

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}$$

with arbitrary noise density N_0 and $\boldsymbol{y}$, $\boldsymbol{x}$ and $\boldsymbol{n}$ are $K\times 1$ vectors. Assume perfect feedback information, i.e., $\hat{\boldsymbol{x}} = \boldsymbol{x}$. The demodulator parameters are taken as $\boldsymbol{V} = \boldsymbol{0}$, $\boldsymbol{R} = -(1+\beta)\boldsymbol{I}$, and $\boldsymbol{G} = \beta\boldsymbol{I}$, β an arbitrary positive real value, then the statistical model for $\hat{\boldsymbol{y}}$ is

$$\hat{\boldsymbol{y}} = \boldsymbol{V}\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{x}} = (1+\beta)\boldsymbol{x}.$$

The GMI in (6) for the pair $(\hat{\boldsymbol{y}}, \boldsymbol{x})$ is

$$I_{\text{GMI}}(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G}) = K\big(1+\log(1+\beta)\big).$$

In order to maximize the GMI, the demodulator will choose $\beta\to\infty$ to make $I_{\text{GMI}}(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G})$ infinite. This is because, except for using the feedback information for interference cancelation, the demodulator uses the prior mean $\hat{\boldsymbol{x}}$ as a signal energy via $\boldsymbol{R}$. A demodulator equipped with these parameters will have significant error propagation and does not have much operational meaning for an iterative receiver. Thus, we can conclude that unless constraints are put on $\boldsymbol{R}$, the GMI value is not relevant.

In this paper we shall investigate three different constraints (to be made precise later) for $\boldsymbol{R}$. All three have in common that rather than adding signal energy, the rationale of $\boldsymbol{R}$ should be to remove interference. Therefore at the very minimum the diagonal elements of $\boldsymbol{R}$ should be constrained to zero, so that the demodulation of each symbol in $\boldsymbol{x}$ does not rely on its own prior mean $\hat{\boldsymbol{x}}$. Such a constraint is perfectly aligned with the operations of the LMMSE-PIC demodulator, where $\hat{x}_\ell$ is not used for demodulation of x_ℓ. Furthermore, the rationale of the constraints we impose on $\boldsymbol{R}$ is to follow the principle of extrinsic information: The BCJR module should not rely on the prior information $\hat{x}_\ell$ when demodulating x_ℓ (this requires more than just the diagonal of $\boldsymbol{R}$ to be zero).

We point out that the fact that the GMI can exceed the channel capacity is a consequence of our choice to not include the side information as a prior distribution on $\boldsymbol{x}$ when evaluating the GMI. If we did, then the GMI is decaying with increasing quality of the side-information.

Finally, we acknowledge the fact that a permutation of the columns of $\boldsymbol{H}$ can boost the performance of the CS demodulator whenever $0 < \nu < K-1$ for finite length linear vector channels and this will be briefly illustrated in the numerical result section. However, minimum-phase conversions of ISI channels are not beneficial as we will anyway solve for the optimal front-end filter $\boldsymbol{V}$.

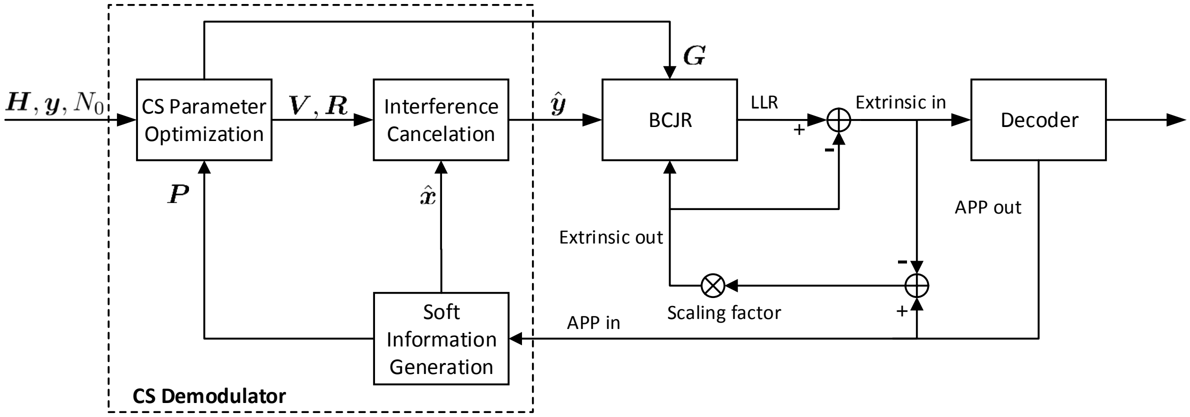

Fig. 2: CS demodulator and decoding receiver structure.

C. Receiver Structure with the CS Demodulator

The overall structure of a receiver that utilizes a CS demodulator is shown in Figure 2. The APP output from the outer decoder is used to compute the estimate $\hat{\boldsymbol{x}}$ and the matrix $\boldsymbol{P}$. Based on the updated $\boldsymbol{P}$ in each global iteration, the optimal CS parameters are found by maximizing the GMI in (6). An interference cancelation process is then implemented with the optimal $\boldsymbol{V}$ and $\boldsymbol{R}$ to obtain the signal $\hat{\boldsymbol{y}}$, which is sent to a memory ν BCJR module with the optimal $\boldsymbol{G}$. Moreover, the extrinsic information iteratively exchanged between the BCJR demodulator and the outer turbo decoder is also used as the priori information for the transmitted symbols.

IV. Parameter Optimization for Finite Length Linear Vector Channel

A. Method I

Method I has its roots in Falconer and Magee's paper [3], but adds an interference cancelation step. The system model of the demodulator is

$$\tilde{T}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}}) = \exp\big(-\|\boldsymbol{W}\boldsymbol{y} - \boldsymbol{T}\hat{\boldsymbol{x}} - \boldsymbol{F}\boldsymbol{x}\|^2\big) \tag{7}$$

where the following structures of the involved matrices are imposed:

- $\boldsymbol{W}$ is a $K \times N$ matrix with no constraints.
- $\boldsymbol{F}$ is a $K \times K$ lower triangular matrix where only the main diagonal and the first ν lower diagonals are non-zero, i.e., $\boldsymbol{F}$ is banded within diagonals $[0, \nu]$ $(0 \leq \nu < K-1)$. ν is denoted as the memory length of $\boldsymbol{F}$. Moreover, the main diagonal of $\boldsymbol{F}$ is constrained to only contain positive real values.

- $\boldsymbol{T}$ is a $K\times K$ matrix that is constrained to be zero wherever $\boldsymbol{F}$ can take non-zero values.

We point out that by setting $\boldsymbol{T}=\boldsymbol{0}$, we obtain the same system model as in [3]. The constraint of $\boldsymbol{F}$ is to shorten the memory for the trellis search in the BCJR module, while the purpose of the constraint on $\boldsymbol{T}$ is to cancel the signal part that $\boldsymbol{F}$ can not handle.

Note that if we identify $\boldsymbol{V}=\boldsymbol{F}^{\mathrm{H}}\boldsymbol{W}$, $\boldsymbol{R}=\boldsymbol{F}^{\mathrm{H}}\boldsymbol{T}$ and $\boldsymbol{G}=\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F}$, (7) can be rewritten in the general form (3),

$$\begin{aligned}\tilde{T}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}}) &\propto \exp\big(2\mathrm{Re}\{\boldsymbol{x}^{\mathrm{H}}(\boldsymbol{F}^{\mathrm{H}}\boldsymbol{W}\boldsymbol{y}-\boldsymbol{F}^{\mathrm{H}}\boldsymbol{T}\hat{\boldsymbol{x}})\}-\boldsymbol{x}^{\mathrm{H}}\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F}\boldsymbol{x}\big)\\ &= \exp\big(2\mathrm{Re}\{\boldsymbol{x}^{\mathrm{H}}(\boldsymbol{V}\boldsymbol{y}-\boldsymbol{R}\hat{\boldsymbol{x}})\}-\boldsymbol{x}^{\mathrm{H}}\boldsymbol{G}\boldsymbol{x}\big)\\ &= \tilde{p}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}}),\end{aligned}$$

and the GMI in (6) in this case reads,

$$\begin{aligned}I_{\mathrm{GMI}}(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F}) = \log\big(\det(\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})\big)-\mathrm{Tr}(\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})+2\mathrm{Re}\big\{\mathrm{Tr}\big(\boldsymbol{F}^{\mathrm{H}}(\boldsymbol{W}\boldsymbol{H}-\boldsymbol{T}\boldsymbol{P})\big)\big\}\\ -\mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})^{-1}\boldsymbol{L}_1\big) \qquad (8)\end{aligned}$$

where

$$\boldsymbol{L}_1=\boldsymbol{F}^{\mathrm{H}}\boldsymbol{W}(N_0\boldsymbol{I}+\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}})\boldsymbol{W}^{\mathrm{H}}\boldsymbol{F}-2\mathrm{Re}\big\{\boldsymbol{F}^{\mathrm{H}}\boldsymbol{W}\boldsymbol{H}\boldsymbol{P}\boldsymbol{T}^{\mathrm{H}}\boldsymbol{F}\big\}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{T}\boldsymbol{P}\boldsymbol{T}^{\mathrm{H}}\boldsymbol{F}.$$

It can be verified that with the aforementioned constraints on $\boldsymbol{F}$ and $\boldsymbol{T}$, the elements of $\boldsymbol{R}=\boldsymbol{F}^{\mathrm{H}}\boldsymbol{T}$ have the special form of type (a) that is depicted in Figure 3. That is, all diagonal elements are zero as well as the lower triangular part of the $(\nu+1)\times(\nu+1)$ right bottom corner.

In order to optimize (8) over the parameters $(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F})$, we first introduce an $S\times K^2$ indication matrix $\boldsymbol{\Omega}$ only consisting of ones and zeros, having a single 1 in each row. S equals the number of elements in $\boldsymbol{T}$ that are allowed to be non-zero. Let $\mathbb{I}(\mathrm{vec}(\boldsymbol{T}))$ be a vector that contains the positions where the vector $\mathrm{vec}(\boldsymbol{T})$ is allowed to be non-zero. Then the value of the kth entry in $\mathbb{I}(\mathrm{vec}(\boldsymbol{T}))$ gives the column where row k of $\boldsymbol{\Omega}$ is 1. That is, the $S\times 1$ vector $\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T})$ stacks the columns of $\boldsymbol{T}$ on top of each other but with all elements that are constrained to zero removed.

With such a definition of $\boldsymbol{\Omega}$, and define two $K\times K$ matrices as,

$$\boldsymbol{M} = \boldsymbol{H}^{\mathrm{H}}(N_0\boldsymbol{I}+\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}})^{-1}\boldsymbol{H}-\boldsymbol{I}, \qquad (9)$$

$$\tilde{\boldsymbol{M}} = \boldsymbol{P}(\boldsymbol{I}+\boldsymbol{M})\boldsymbol{P}-\boldsymbol{P}, \qquad (10)$$

the GMI for the optimal $\boldsymbol{W}$ and $\boldsymbol{T}$ is given in Proposition 1 and the proof is in Appendix B.

Proposition 1. *Define an $S\times K^2$ matrix $\boldsymbol{D}=\boldsymbol{\Omega}\big((\boldsymbol{P}\boldsymbol{M}^*)\otimes\boldsymbol{I}_K\big)$, the optimal $\boldsymbol{W}$ for the GMI in (8) is,*

$$\boldsymbol{W}_{\text{opt}}=\boldsymbol{F}^{-\text{H}}(\boldsymbol{I}+\boldsymbol{F}^{\text{H}}\boldsymbol{F}+\boldsymbol{F}^{\text{H}}\boldsymbol{T}\boldsymbol{P})\boldsymbol{H}^{\text{H}}(N_0\boldsymbol{I}+\boldsymbol{H}\boldsymbol{H}^{\text{H}})^{-1}, \tag{11}$$

and when $\boldsymbol{P}\neq\boldsymbol{0}$, the optimal $\boldsymbol{T}$ for the GMI in (8) is given by,

$$\text{vec}(\boldsymbol{T}_{\text{opt}})=-\boldsymbol{\Omega}^{\text{T}}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\text{H}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\text{H}}\big)\big)\boldsymbol{\Omega}^{\text{T}}\big)^{-1}\boldsymbol{D}\text{vec}(\boldsymbol{F}). \tag{12}$$

With the optimal $\boldsymbol{W}$ and $\boldsymbol{T}$, the GMI reads,

$$I_{\text{GMI}}(\boldsymbol{W}_{\text{opt}},\boldsymbol{T}_{\text{opt}},\boldsymbol{F})=\begin{cases}I_1(\boldsymbol{F}), & \boldsymbol{P}=\boldsymbol{0}\\ I_1(\boldsymbol{F})+\delta_1(\boldsymbol{F}), & \boldsymbol{P}\neq\boldsymbol{0}.\end{cases} \tag{13}$$

The functions $I_1(\boldsymbol{F})$ and $\delta_1(\boldsymbol{F})$[3] are defined as,

$$I_1(\boldsymbol{F})=K+\log\big(\det(\boldsymbol{I}+\boldsymbol{F}^{\text{H}}\boldsymbol{F})\big)+\text{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{F}^{\text{H}}\boldsymbol{F})\big), \tag{14}$$

$$\delta_1(\boldsymbol{F})=-\text{vec}(\boldsymbol{F})^{\text{H}}\boldsymbol{D}^{\text{H}}\Big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\text{H}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\text{H}}\big)\big)\boldsymbol{\Omega}^{\text{T}}\Big)^{-1}\boldsymbol{D}\text{vec}(\boldsymbol{F}). \tag{15}$$

Remark 1. *With the definitions in (9) and (10), $\boldsymbol{M}$ is the negative of the MSE matrix. By the matrix inversion lemma [34], $\boldsymbol{M}=-(\boldsymbol{I}+\boldsymbol{H}^{\text{H}}\boldsymbol{H}/N_0)^{-1}\prec 0$ and $\tilde{\boldsymbol{M}}=\boldsymbol{P}\boldsymbol{M}\boldsymbol{P}+\boldsymbol{P}^2-\boldsymbol{P}\prec 0$. Hence $\delta_1(\boldsymbol{F})>0$ and it represents the GMI increment from the soft information feedback.*

Before discussing the optimization of (13), we state Theorem 2 that deals with a general GMI maximization problem.

Theorem 2. *Define a scalar function I with respect to a $K\times K$ matrix $\boldsymbol{G}$ as*

$$I(\boldsymbol{G})=K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+\text{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big), \tag{16}$$

where $\boldsymbol{G}$ satisfies $\boldsymbol{G}=[\boldsymbol{G}]_\nu$. Then the optimal $\boldsymbol{G}$ that maximizes I is the unique solution that satisfies

$$[(\boldsymbol{I}+\boldsymbol{G}_{\text{opt}})^{-1}]_\nu=-[\boldsymbol{M}]_\nu. \tag{17}$$

With $\boldsymbol{G}_{\text{opt}}$ the optimal I reads,

$$I(\boldsymbol{G}_{\text{opt}})=\log\big(\det(\boldsymbol{I}+\boldsymbol{G}_{\text{opt}})\big). \tag{18}$$

[3] $\delta_1(\boldsymbol{F})$ in (15) is only defined for $\boldsymbol{P}\neq\boldsymbol{0}$, as when $\boldsymbol{P}=\boldsymbol{0}$, $\tilde{\boldsymbol{M}}=\boldsymbol{0}$ and the inversion in $\delta_1(\boldsymbol{F})$ is not well defined. The same comment holds for $\delta_2(\boldsymbol{G})$ in (25).

Proof: Taking the first order differential of I with respect to $\boldsymbol{G}$ and noticing that $\boldsymbol{G}$ is banded within diagonals $[-\nu,\nu]$, yields (17) after some manipulations. The existence and uniqueness of such an optimal solution for (17) is proved in [35, Theorem 2] and also illustrated in [18, Proposition 2]. By Lemma 1, $\mathrm{Tr}\big([\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}}]^{-1}\boldsymbol{M}\big)=-K$ from (17), and then (18) follows. ■

Optimizing over $\boldsymbol{F}$ in (13) when $\boldsymbol{P}\neq\boldsymbol{0}$ is difficult and cannot be carried out in closed form. In Appendix C we show by an example that (13) is in general non-concave. Nevertheless, a gradient based numerical optimization procedure is utilized to search for the optimal $\boldsymbol{F}$. In the ith iteration, we construct

$$\boldsymbol{F}^{(i)}=\boldsymbol{F}^{(i-1)}+\nabla_{\boldsymbol{F}^*}I_{\mathrm{GMI}}\big(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T}_{\mathrm{opt}},\boldsymbol{F}^{(i-1)}\big)$$

where $\nabla_{\boldsymbol{F}^*}I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T}_{\mathrm{opt}},\boldsymbol{F})$ is the conjugate of the gradient $\nabla_{\boldsymbol{F}}I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T}_{\mathrm{opt}},\boldsymbol{F})$ with respect to (the non-zero part of) $\boldsymbol{F}$, which is given in Appendix D.

When $\boldsymbol{P}=\boldsymbol{0}$, if we replace $\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F}$ by $\boldsymbol{G}$, (14) has the same form as (16) in Theorem 2 and $\boldsymbol{G}_{\mathrm{opt}}$ is in closed form. However, it can be verified that (14) is non-concave with respect to $\boldsymbol{F}$. Hence, if the optimal $\boldsymbol{G}$ is positive definite, the optimal $\boldsymbol{F}$ equals the Cholesky decomposition of the optimal $\boldsymbol{G}$. Whenever it is not, a gradient based numerical optimization procedure is utilized to optimize (14). After applying a regularization to force the optimal $\boldsymbol{G}$ to be positive definite, the Cholesky decomposition of the optimal $\boldsymbol{G}$ is chosen to be the starting point of $\boldsymbol{F}$ in the optimization procedure both for the case $\boldsymbol{P}\neq\boldsymbol{0}$ and $\boldsymbol{P}=\boldsymbol{0}$. The optimization procedure has been observed to be highly reliable with such initialization.

Next we state a fact that establishes the connection between the optimal front-end filter $\boldsymbol{W}$ and the optimal interference cancelation matrix $\boldsymbol{T}$ in Proposition 2.

Proposition 2. *For $\boldsymbol{P}\neq\boldsymbol{0}$ and the optimal $\boldsymbol{W}$ and $\boldsymbol{T}$, the matrix $\boldsymbol{F}^{\mathrm{H}}(\boldsymbol{W}_{\mathrm{opt}}\boldsymbol{H}-\boldsymbol{R}_{\mathrm{opt}})$ is banded within diagonals $[-\nu,K-1]$ for any $\boldsymbol{F}$ that is banded within diagonals $[0,\nu]$.*

Proof: By the definition of $\boldsymbol{\Omega}$, we have $\boldsymbol{\Omega}^{\mathrm{T}}\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T}_{\mathrm{opt}})=\mathrm{vec}(\boldsymbol{T}_{\mathrm{opt}})$ and $\boldsymbol{\Omega}\boldsymbol{\Omega}^{\mathrm{T}}=\boldsymbol{I}$. Therefore (12) can be rewritten as,

$$\begin{aligned}\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*}\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{H}}\big)\big)\boldsymbol{\Omega}^{\mathrm{T}}\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T}_{\mathrm{opt}}) &= \boldsymbol{\Omega}\mathrm{vec}\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{H}}\boldsymbol{T}_{\mathrm{opt}}\tilde{\boldsymbol{M}}\big)\\ &= -\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{F}\boldsymbol{M}\boldsymbol{P}).\end{aligned}\qquad(19)$$

This shows that the elements of the matrix $\Delta=\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{H}}\boldsymbol{T}_{\mathrm{opt}}\tilde{\boldsymbol{M}}+\boldsymbol{F}\boldsymbol{M}\boldsymbol{P}$ are zero wherever $\boldsymbol{T}$ can be non-zero. Hence Δ is banded within diagonals $[0,\nu]$. On the other hand, with the optimal $\boldsymbol{W}$ given in (11) and $\boldsymbol{M}$, $\tilde{\boldsymbol{M}}$ defined in (9) and (10), we have

$$\boldsymbol{F}^{\mathrm{H}}(\boldsymbol{W}_{\mathrm{opt}}\boldsymbol{H}-\boldsymbol{T}_{\mathrm{opt}})-(\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F}) = (\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})\boldsymbol{F}^{-1}\Delta\boldsymbol{P}^{-1}. \tag{20}$$

Note that $\boldsymbol{F}^{-1}$ is lower triangular since $\boldsymbol{F}$ is lower triangular, $\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F}$ is banded within diagonals $[-\nu,\nu]$, and $\boldsymbol{P}$ is diagonal. Utilizing Lemma 2, the r.h.s in (20) is banded within diagonals $[-\nu, K-1]$. Therefore $\boldsymbol{F}^{\mathrm{H}}(\boldsymbol{W}_{\mathrm{opt}}\boldsymbol{H}-\boldsymbol{T}_{\mathrm{opt}})$ is also banded within diagonals $[-\nu, K-1]$. ■

B. Method II

Method II origins from Ungerboeck's 1974 paper [33]. Different from Method I, an Ungerboeck detection model (3) instead of the Forney model (7) is applied. The Ungerboeck model has been extensively discussed in [36]–[38]. The system model (3) in Method II has the following constraints for the involved matrices:

- $\boldsymbol{V}$ is a $K\times N$ matrix with no constraints.
- $\boldsymbol{G}$ is a $K\times K$ Hermitian matrix that satisfies $\boldsymbol{G}=[\boldsymbol{G}]_{\nu}$ and $\boldsymbol{G}+\boldsymbol{I}\succ 0$. ν is denoted as the memory length of $\boldsymbol{G}$.
- $\boldsymbol{R}$ is a $K\times K$ matrix where the shape can be specified and three typical shapes are of our interest and investigated.

Instead of optimizing matrices $(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F})$ for (8) in Method I, we now optimize matrices $(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G})$ for (6) directly. In Method II we use the same definition of the indication matrix $\boldsymbol{\Omega}$ as in Method I, but now $\boldsymbol{\Omega}$ corresponds to matrix $\boldsymbol{R}$ instead of $\boldsymbol{T}$. We continue to let S denote the number of elements that are allowed to be non-zero in $\boldsymbol{R}$. That is, the $S\times 1$ vector $\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{R})$ stacks the columns of $\boldsymbol{R}$ on top of each other but with all elements that are constrained to zero removed. Then we have the following Proposition 3.

Proposition 3. *Define an $S\times 1$ vector $\boldsymbol{d}=\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{M}\boldsymbol{P})$, the optimal $\boldsymbol{V}$ for the GMI in (6) is,*

$$\boldsymbol{V}_{\mathrm{opt}} = (\boldsymbol{I}+\boldsymbol{G}+\boldsymbol{R}_{\mathbf{opt}}\boldsymbol{P})\boldsymbol{H}^{\mathrm{H}}(\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}}+N_0\boldsymbol{I})^{-1}, \tag{21}$$

and when $\boldsymbol{P}\neq\boldsymbol{0}$, the optimal $\boldsymbol{R}$ for the GMI in (6) is given by,

$$\mathrm{vec}(\boldsymbol{R}_{\mathrm{opt}}) = -\boldsymbol{\Omega}^{\mathrm{T}}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*}\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}\big)^{-1}\boldsymbol{d}. \tag{22}$$

With the optimal $\boldsymbol{V}$ and $\boldsymbol{R}$, the GMI in (6) equals

$$I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R}_{\mathrm{opt}},\boldsymbol{G}) = \begin{cases} I_2(\boldsymbol{G}), & \boldsymbol{P}=\boldsymbol{0} \\ I_2(\boldsymbol{G})+\delta_2(\boldsymbol{G}), & \boldsymbol{P}\neq\boldsymbol{0}. \end{cases} \tag{23}$$

The functions $I_2(\boldsymbol{G})$ and $\delta_2(\boldsymbol{G})$ are defined as,

$$I_2(\boldsymbol{G}) = K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big), \tag{24}$$

$$\delta_2(\boldsymbol{G}) = -\boldsymbol{d}^{\mathrm{H}}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*}\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}\big)^{-1}\boldsymbol{d}. \tag{25}$$

The proof is given in Appendix E. Similar to $\delta_1(\boldsymbol{F})$ in Method I, $\delta_2(\boldsymbol{G})>0$ represents the GMI increment from the soft information feedback in Method II.

When $\boldsymbol{P}\neq\boldsymbol{0}$, the optimization over $\boldsymbol{G}$ in (23) uses a gradient based numerical optimization procedure and the gradient of $I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R}_{\mathrm{opt}},\boldsymbol{G})$ with respect to (the non-zero part of) $\boldsymbol{G}$ is provided in Appendix F. When $\boldsymbol{P}=\boldsymbol{0}$, the optimal $\boldsymbol{G}$ for (24) is provided in closed form Theorem 2 and used as the starting point for the optimization procedure for $\boldsymbol{P}\neq\boldsymbol{0}$. However, different from Method I, the optimization procedure is concave and the proof is given in Appendix G.

Although the optimal matrix $\boldsymbol{R}$ is solved for in closed form as in (22), we have not specified the constraint (reflected by the indication matrix $\boldsymbol{\Omega}$) on $\boldsymbol{R}$ yet. As we are interested in the comparison between Method I and Method II, we introduce a first shape, type (a), that is the same as for $\boldsymbol{R}=\boldsymbol{F}^{\mathrm{H}}\boldsymbol{T}$ in Method I. The second shape, type (b), is that we only limit the diagonal elements of $\boldsymbol{R}$ to be zero, the reason is that we intended to eliminate the interference as well as possible. At last we introduce shape type (c), in which we limit $\boldsymbol{R}$ to have the opposite form of matrix $\boldsymbol{G}$, that is, the elements of $\boldsymbol{R}$ are constrained to be zero wherever $\boldsymbol{G}$ is non-zero. The intention is to only cancel the interference that the trellis search process in BCJR represented by $\boldsymbol{G}$ cannot handle. Shape (c) is based on the same idea as Method I, but operates on the Ungerboeck model instead of the Forney model. These three types are depicted in Figure 3 and ν is the memory length constraint for $\boldsymbol{G}$. In the following we refer to "Method II with an $\boldsymbol{R}$ of shape type (a), type (b) and type (c)" as "Method II.a", "Method II.b", and "Method II.c," respectively.

Similar to Method I, the connections between the optimal front-end filter $\boldsymbol{V}$ and the optimal interference cancelation matrix $\boldsymbol{R}$ in Method II are established in Proposition 4.

Proposition 4. *For $\boldsymbol{P}\neq 0$ and the optimal $\boldsymbol{V}$ and $\boldsymbol{R}$,*

$$[\boldsymbol{V}_{\mathrm{opt}}\boldsymbol{H}]_{\backslash(\nu+\nu_{\mathrm{R}})} = [\boldsymbol{R}_{\mathrm{opt}}]_{\backslash(\nu+\nu_{\mathrm{R}})}. \tag{26}$$

That is, the elements of $\boldsymbol{V}_{\mathrm{opt}}\boldsymbol{H}$ and $\boldsymbol{R}_{\mathrm{opt}}$ are equal outside the center $2(\nu+\nu_{\mathrm{R}})+1$ diagonals for

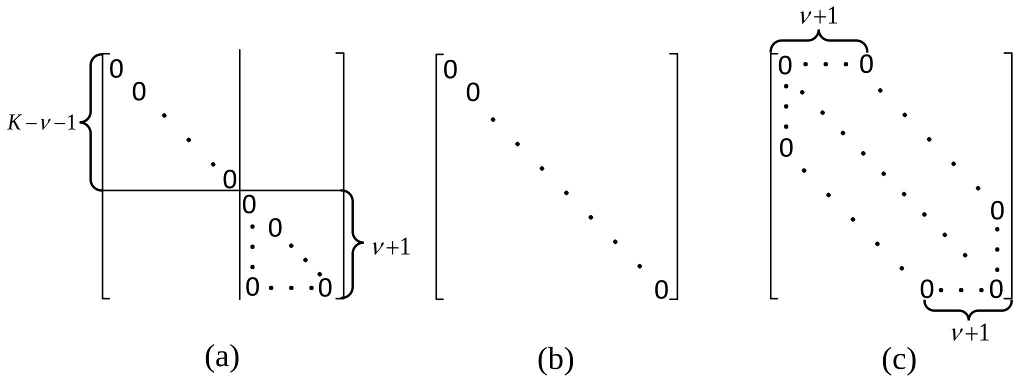

Fig. 3: Three different types of shape of matrix $\boldsymbol{R}$.

any $\boldsymbol{G}$ that is banded within diagonals $[-\nu,\nu]$, where for Method II.a and Method II.b, $\nu_{\mathrm{R}}=0$ and for Method II.c, $\nu_{\mathrm{R}}=\nu$.

Proof: Following similar steps as in the proof of Proposition 2, (22) can be rewritten as,

$$\boldsymbol{\Omega}\mathrm{vec}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\boldsymbol{R}_{\mathrm{opt}}\tilde{\boldsymbol{M}}\big) = -\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{M}\boldsymbol{P}). \tag{27}$$

It shows that the elements of the matrix $\Delta=(\boldsymbol{I}+\boldsymbol{G})^{-1}\boldsymbol{R}_{\mathrm{opt}}\tilde{\boldsymbol{M}}\boldsymbol{P}^{-1}+\boldsymbol{M}$ are zero wherever $\boldsymbol{R}$ can be non-zero. On the other hand, with the optimal $\boldsymbol{V}$ given in (21) we have

$$\boldsymbol{V}_{\mathrm{opt}}\boldsymbol{H}-\boldsymbol{R}_{\mathrm{opt}}-(\boldsymbol{I}+\boldsymbol{G})=(\boldsymbol{I}+\boldsymbol{G})\Delta. \tag{28}$$

As $\boldsymbol{I}+\boldsymbol{G}$ is banded within diagonals $[-\nu,\nu]$, utilizing Lemma 2 (the type (a) of $\boldsymbol{R}$ is sligtly different, but it can be verified straightforwardly), with the three types of $\boldsymbol{R}$ defined in Figure 3, it can be shown that the r.h.s in (28) is banded within diagonals $[-(\nu+\nu_{\mathrm{R}}),\nu+\nu_{\mathrm{R}}]$, where $\nu_{\mathrm{R}}=0$ for the type (a) and type (b) of $\boldsymbol{R}$ and $\nu_{\mathrm{R}}=\nu$ fot the type (c) of $\boldsymbol{R}$. Therefore $\boldsymbol{V}_{\mathrm{opt}}\boldsymbol{H}-\boldsymbol{R}_{\mathrm{opt}}$ on the l.h.s in (28) is banded within diagonals $[-(\nu+\nu_{\mathrm{R}}),\nu+\nu_{\mathrm{R}}]$, which proves Proposition 3 ■

Remark 2. *With the LMMSE-PIC demodulator, we have $\nu=\nu_{\mathrm{R}}=0$ and Proposition 4 is natural and frequently used. With CS demodulators and $\nu\neq 0$, $\boldsymbol{V}_{\mathrm{opt}}\boldsymbol{H}$ and $\boldsymbol{R}$ are equal outside the center $2(\nu+\nu_{\mathrm{R}})+1$ diagonals, not the center $2\nu+1$ diagonals where $\boldsymbol{G}$ are constrained to be non-zero. This reveals an interesting fact that the signal part that is not considered in $\boldsymbol{G}$ shall not be perfectly canceled inside the center 2ν+$2\nu_{\mathrm{R}}$+1 diagonals. The LMMSE-PIC demodulator follows this law, but the full nature of the interference cancelation process given in Proposition 4 is not seen with LMMSE-PIC as $\nu=\nu_{\mathrm{R}}=0$.*

C. Method III

So far we have discussed two types of CS demodulators which are based on the Forney and Ungerboeck models. However, as both need numerical optimization to obtain the optimal CS parameter $\boldsymbol{F}$ or $\boldsymbol{G}$, we provide a third method that has a closed form solution but is suboptimal in general. Method III will rely on the same operations as Method II for $\boldsymbol{P}=\boldsymbol{0}$.

With $\boldsymbol{P}=\boldsymbol{0}$, which means that no soft information of the transmitted signal $\boldsymbol{x}$ is available, the GMI is given in (24). The optimal $\boldsymbol{G}$ can be derived from $\boldsymbol{M}$ from Theorem 2 following [18, Proposition 2]. By inserting $\boldsymbol{V}_{\rm opt}$ given in (21) into (3) and setting $\boldsymbol{R}=\boldsymbol{0}$, we can see that the demodulator operates on the mismatched function

$$\begin{aligned}\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) &= \exp\big(2\mathrm{Re}\{\boldsymbol{x}^{\rm H}\boldsymbol{V}_{\rm opt}\boldsymbol{y}\}-\boldsymbol{x}^{\rm H}\boldsymbol{G}\boldsymbol{x}\big)\\ &= \exp\big(2\mathrm{Re}\big\{\boldsymbol{x}^{\rm H}(\boldsymbol{I}+\boldsymbol{G})\boldsymbol{H}^{\rm H}(\boldsymbol{H}\boldsymbol{H}^{\rm H}+N_0\boldsymbol{I})^{-1}\boldsymbol{y}\big\}-\boldsymbol{x}^{\rm H}\boldsymbol{G}\boldsymbol{x}\big)\\ &= \exp\big(2\mathrm{Re}\big\{\boldsymbol{x}^{\rm H}(\boldsymbol{I}+\boldsymbol{G})\check{\boldsymbol{x}}\big\}-\boldsymbol{x}^{\rm H}\boldsymbol{G}\boldsymbol{x}\big) \qquad (29)\end{aligned}$$

where $\check{\boldsymbol{x}}=\boldsymbol{H}^{\rm H}(\boldsymbol{H}\boldsymbol{H}^{\rm H}+N_0\boldsymbol{I})^{-1}\boldsymbol{y}$ is the LMMSE estimate.

As can be seen in (29), the trellis search is based on $\check{\boldsymbol{x}}$. With soft information we can replace the LMMSE estimate $\check{\boldsymbol{x}}$ by the LMMSE-PIC estimate which we denote as $\tilde{\boldsymbol{x}}$. That is, instead of (29) we now operate on the mismatched function

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}}) = \exp\big(2\mathrm{Re}\big\{\boldsymbol{x}^{\rm H}(\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{x}}\big\}-\boldsymbol{x}^{\rm H}\boldsymbol{G}\boldsymbol{x}\big) \qquad (30)$$

where $\boldsymbol{G}$ has the same banded shape as in Method II, i.e., $\boldsymbol{G}=[\boldsymbol{G}]_\nu$.

The LMMSE-PIC estimate $\tilde{\boldsymbol{x}}$ is constructed as follows. As we prefer to handle the interference through the trellis search process, the interference cancelation should not be present within the memory length constraint ν. In other words, the signal vector after the interference cancelation process that is used to form the kth symbol of $\tilde{\boldsymbol{x}}$ is denoted as $\tilde{\boldsymbol{y}}_k$ and defined as

$$\tilde{\boldsymbol{y}}_k=\boldsymbol{y}-\sum_{n\in\mathcal{A}_k}\boldsymbol{h}_n\hat{x}_n \qquad (31)$$

where $\mathcal{A}_k=\big\{0\leq n\leq K-1 : n\notin[\max(0,k-\nu),\min(k+\nu,K-1)]\big\}$.

Denote p_n as the nth diagonal element of $\boldsymbol{P}$, the Wiener filtering coefficients [39] for the kth symbol are calculated through

$$\hat{\boldsymbol{w}}_k = \boldsymbol{h}_k^{\rm H}\big(\boldsymbol{H}^{\rm H}\boldsymbol{C}_k\boldsymbol{H}+N_0\boldsymbol{I}\big)^{-1} \qquad (32)$$

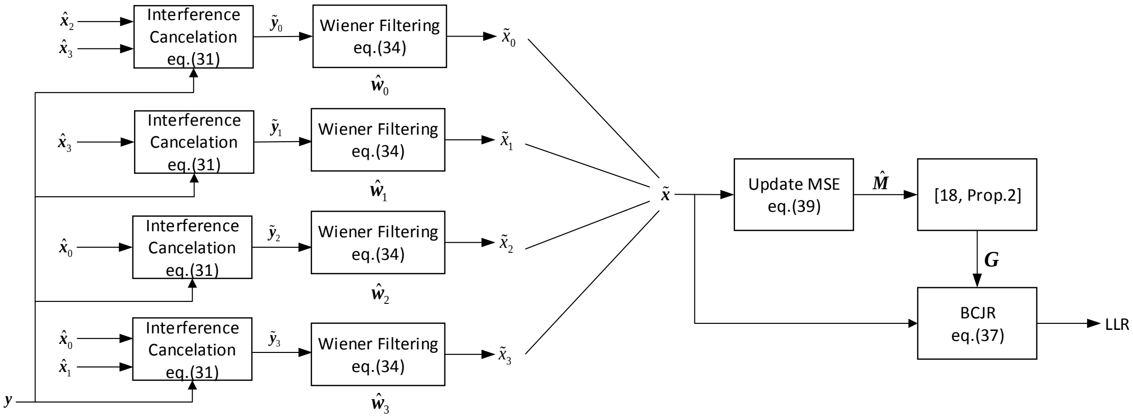

Fig. 4: An graphical overview of Method III with $K\!=\!4$ and $\nu\!=\!1$.

where $\boldsymbol{C}_k$ is a diagonal matrix with the nth diagonal element defined as

$$C_k(n) = \begin{cases} 1-p_n, & k\in\mathcal{A}_k \\ 1, & \text{otherwise.} \end{cases} \tag{33}$$

The LMMSE-PIC estimate $\tilde{\boldsymbol{x}}$ is then obtained through

$$\tilde{\boldsymbol{x}} = \big[\hat{\boldsymbol{w}}_1\tilde{\boldsymbol{y}}_1 \quad \hat{\boldsymbol{w}}_2\tilde{\boldsymbol{y}}_2 \quad \cdots \quad \hat{\boldsymbol{w}}_K\tilde{\boldsymbol{y}}_K\big]^{\mathrm{T}} = \hat{\boldsymbol{W}}\boldsymbol{y} - \hat{\boldsymbol{C}}\hat{\boldsymbol{x}} \tag{34}$$

with the coefficient matrix $\hat{\boldsymbol{W}}$ and interference cancelation matrix $\hat{\boldsymbol{C}}$ defined as,

$$\hat{\boldsymbol{W}} = [\,\hat{\boldsymbol{w}}_1^{\mathrm{T}}\ \hat{\boldsymbol{w}}_2^{\mathrm{T}}\ \cdots\ \hat{\boldsymbol{w}}_K^{\mathrm{T}}\,]^{\mathrm{T}}, \tag{35}$$

$$\hat{\boldsymbol{C}} = [\hat{\boldsymbol{W}}\boldsymbol{H}]_{\backslash\nu}. \tag{36}$$

Putting $\tilde{\boldsymbol{x}}$ in (34) back into (30), the system model we operate on reads,

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}}) = \exp\!\big(2\mathrm{Re}\big\{\boldsymbol{x}^{\mathrm{H}}\big((\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{W}}\boldsymbol{y} - (\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{C}}\hat{\boldsymbol{x}}\big)\big\} - \boldsymbol{x}^{\mathrm{H}}\boldsymbol{G}\boldsymbol{x}\big). \tag{37}$$

Note that (37) is a also special case of (3) by identifying the front-end filter $\boldsymbol{V} = (\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{W}}$ and the interference cancelation matrix $\boldsymbol{R}\!=\!(\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{C}}$. The GMI in (6) in this case reads, after some manipulations,

$$I_{\mathrm{GMI}}(\boldsymbol{G}) = K + \log\!\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) + \mathrm{Tr}\big(\hat{\boldsymbol{M}}(\boldsymbol{I}+\boldsymbol{G})\big) \tag{38}$$

with $\hat{\boldsymbol{M}}$ defined as

$$\begin{aligned} \hat{\boldsymbol{M}} = {} & \hat{\boldsymbol{W}}\boldsymbol{H}\boldsymbol{P}\hat{\boldsymbol{C}}^{\mathrm{H}} + \hat{\boldsymbol{W}}\boldsymbol{H} - \boldsymbol{P}\hat{\boldsymbol{C}}^{\mathrm{H}} + \big(\hat{\boldsymbol{W}}\boldsymbol{H}\boldsymbol{P}\hat{\boldsymbol{C}}^{\mathrm{H}} + \hat{\boldsymbol{W}}\boldsymbol{H} - \boldsymbol{P}\hat{\boldsymbol{C}}^{\mathrm{H}}\big)^{\mathrm{H}} \\ & - \hat{\boldsymbol{W}}(\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}} + N_0\boldsymbol{I})\hat{\boldsymbol{W}}^{\mathrm{H}} - \hat{\boldsymbol{C}}\boldsymbol{P}\hat{\boldsymbol{C}}^{\mathrm{H}} - \boldsymbol{I}. \end{aligned} \tag{39}$$

It can be verified that $\hat{\boldsymbol{M}}$ is the negative of the updated MSE matrix, that is,

$$\hat{\boldsymbol{M}} = -\mathbb{E}\big[(\boldsymbol{x}-\tilde{\boldsymbol{x}})(\boldsymbol{x}-\tilde{\boldsymbol{x}})^{\mathrm{H}}\big] = -\mathbb{E}\big[(\boldsymbol{x}-\hat{\boldsymbol{W}}\boldsymbol{y}+\hat{\boldsymbol{C}}\hat{\boldsymbol{x}})(\boldsymbol{x}-\hat{\boldsymbol{W}}\boldsymbol{y}+\hat{\boldsymbol{C}}\hat{\boldsymbol{x}})^{\mathrm{H}}\big].$$

The optimal $\boldsymbol{G}$ for (38) is obtained from Theorem 2 and the optimal GMI is

$$I_{\mathrm{GMI}}(\boldsymbol{G}_{\mathrm{opt}}) = \log\big(\det(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})\big).$$

An graphical overview of Method III for $K=4$ and $\nu=1$ is illustrated in Figure 4. Note that for any band shaped matrix $\boldsymbol{G}$ with memory length ν, the interference cancelation matrix $(\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{C}}$ is zero along the main diagonal. Therefore in GMI sense, Method III will not outperform Method II.b. But the GMI of Method III may outperform the GMI of Method II.c, as it can be verified that a type (c) $\boldsymbol{R}$ has zeros at the positions where $(\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{C}}$ is zero, therefore Method II.c has less degrees of freedom (DoFs) for designing $\boldsymbol{R}$ than Method III.

Remark 3. *Proposition 4 also holds for Method III with $\nu_{\mathrm{R}} = \nu$. As $\hat{\boldsymbol{W}}\boldsymbol{H} - \hat{\boldsymbol{C}} = [\hat{\boldsymbol{W}}\boldsymbol{H}]_{\nu}$, by Lemma 2 $(\boldsymbol{I}+\boldsymbol{G})(\hat{\boldsymbol{W}}\boldsymbol{H} - \hat{\boldsymbol{C}})$ is banded within diagonals $[-2\nu, 2\nu]$, which shows that, $[(\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{W}}\boldsymbol{H}]_{\backslash 2\nu} = [(\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{C}}]_{\backslash 2\nu}$.*

V. Parameter Optimization for ISI Channel

In this section, we extend the CS demodulators to ISI channels for all three methods. The difference from finite length linear vector channels is that with ISI channels the channel matrix is infinitely large, but this can be dealt with using [40]–[42]. The formulas for the achievable rates in (6), (8) and (38) can be directly applied to (1), but as the achievable rate I_{GMI} (as a function of the specified CS parameters) is then dependent on the block length K, we are interested in the asymptotic rate

$$\bar{I} = \lim_{K\to\infty} \frac{1}{K} I_{\mathrm{GMI}}.$$

Ideally, in the ISI case the front-end matrices $\boldsymbol{W}$ and $\boldsymbol{V}$ correspond to linear filtering operations. The filters are infinitely long, but in practice filters with finite tap lengths are used. Therefore, we analyze the properties of $\boldsymbol{W}$ and $\boldsymbol{V}$ with a finite number of taps. In other words, we approximate $\boldsymbol{W}$ and $\boldsymbol{V}$ by band shaped matrices and constrain $\boldsymbol{W}$ and $\boldsymbol{V}$ to be zero outside the band. However, the band size can be arbitrary and sufficiently large so that we can analyze the asymptotic properties. The same holds for the interference cancelation matrices $\boldsymbol{T}$ and $\boldsymbol{R}$, as they are Toeplitz matrices in the ISI case and multiplying $\boldsymbol{T}$ and $\boldsymbol{R}$ with $\hat{\boldsymbol{x}}$ can be replaced by filtering operations. Therefore, $\boldsymbol{T}$ and $\boldsymbol{R}$ are also approximated by band shaped matrices. Moreover, the trellis representation matrices $\boldsymbol{F}$ and $\boldsymbol{G}$ are by definition constrained to be band shaped with a limited memory length ν, and the channel matrix $\boldsymbol{H}$ itself is band shaped Toeplitz matrix. Therefore, in the ISI case all matrices we consider are assumed to be band shaped Toeplitz matrices. In [40] a complete theoretic machinery for ISI channels is derived, and a result is that as $K\rightarrow\infty$, the linear convolution in (1) can be replaced with a circular convolution.[4] We can then let $\boldsymbol{H}$ and all other band shaped Toeplitz matrices represent circular convolutional matrices, and apply Szegö's eigenvalue distribution theorem [41] in order to evaluate $\bar{I}$.

In the following, we denote the Fourier series associated to a band shaped Toeplitz matrix $\boldsymbol{E}$ that has infinitely large dimensions by $E(\omega)$. $\boldsymbol{E}$ is constrained to be zero except the middle $2N_{\mathrm{E}}+1$ diagonals, i.e., the band size is $2N_{\mathrm{E}}+1$. N_{E} is referred to as the tap length for $E(\omega)$. The Fourier series $E(\omega)$ is specified by the vector $\boldsymbol{e}=[\,e_{-N_{\mathrm{E}}}\;\ldots\;e_{-1}\;e_0\;e_1\;\ldots\;e_{N_{\mathrm{E}}}\,]$, where e_0 is the element on the main diagonal, e_k $(k>0)$ is the element on kth lower diagonal, and e_{-k} is the element on kth upper diagonal of $\boldsymbol{E}$. The Fourier series $E(\omega)$ is defined as [42]

$$E(\omega)=\sum_{k=-N_{\mathrm{E}}}^{N_{\mathrm{E}}} e_k \exp(jk\omega)\,.$$

As all quantities are evaluated as the block length grows large, the transform $E(\omega)$ approaches the eigenvalue distribution of $\boldsymbol{E}$ (see [41], [42] for a precise statement of this result). The element e_k can be obtained from $E(\omega)$ through an ordinary inverse Fourier formula.

Furthermore, since the whole data block experiences the same channel, we assume $\boldsymbol{P}=\alpha\boldsymbol{I}$ $(0\leq\alpha\leq 1)$, which refers to the quality of the side information. We first state Theorem 3, which is an asymptotic version of Theorem 2 with ISI channels.

Theorem 3. *Assume that $G(\omega)$ and $M(\omega)$ are the Fourier series associated to band shaped Toeplitz matrix $\boldsymbol{G}$ and $\boldsymbol{M}$, respectively. The dimensions of $\boldsymbol{G}$ and $\boldsymbol{M}$ are infinitely large and $\boldsymbol{G}$ is constrained to be zero outside the center $2\nu+1$ diagonals. Moreover, we assume $\boldsymbol{I}+\boldsymbol{G}\succ 0$*

[4] A conceptually simple way to realize this is to insert a cyclic prefix (of ISI channel tap length L) and then make the observation that the cyclic prefix has vanishing impact on energy and spectral-efficiency as $K\rightarrow\infty$.

and $\boldsymbol{M}\prec 0$. *Define a scalar function* $\bar{I}$ *with respect to* $G(\omega)$ *as*

$$\bar{I}(G(\omega)) = 1+\frac{1}{2\pi}\int_{-\pi}^{\pi}\big(\log(1+G(\omega))+M(\omega)(1+G(\omega))\big)\mathrm{d}\omega, \tag{40}$$

and a $1\times\nu$ *vector*

$$\varphi(\omega)=[\,\exp(j\omega)\ \exp(j2\omega)\,\ldots\,\exp(j\nu\omega)\,]^{\mathrm{T}},$$

then the optimal $G(\omega)$ *that maximizes* $\bar{I}$ *in (40) is*

$$G(\omega)_{\mathrm{opt}}=|u_0+\hat{\boldsymbol{u}}\varphi(\omega)|^2-1,$$

where

$$\begin{aligned} u_0 &= \frac{1}{\sqrt{\boldsymbol{\tau}_1^{\mathrm{H}}\boldsymbol{\tau}_2^{-1}\boldsymbol{\tau}_1-\tau_0}}, \\ \hat{\boldsymbol{u}} &= -u_0\boldsymbol{\tau}_1^{\mathrm{H}}\boldsymbol{\tau}_2^{-1}, \end{aligned} \tag{41}$$

and the real scalar τ_0*,* $\nu\times 1$ *vector* $\boldsymbol{\tau}_1$*, and* $\nu\times\nu$ *matrix* $\boldsymbol{\tau}_2$ *are defined as*

$$\begin{aligned} \tau_0 &= \frac{1}{2\pi}\int_{-\pi}^{\pi}M(\omega)\mathrm{d}\omega, \\ \boldsymbol{\tau}_1 &= \frac{1}{2\pi}\int_{-\pi}^{\pi}M(\omega)\varphi(\omega)\mathrm{d}\omega, \\ \boldsymbol{\tau}_2 &= \frac{1}{2\pi}\int_{-\pi}^{\pi}M(\omega)\varphi(\omega)\varphi(\omega)^{\mathrm{H}}\mathrm{d}\omega. \end{aligned}$$

Furthermore, the optimal $\bar{I}$ *is*

$$\bar{I}(G_{\mathrm{opt}}(\omega)) = 2\log(u_0). \tag{42}$$

Proof: As $\boldsymbol{I}+\boldsymbol{G}\succ 0$, in order to maximize (40), we assume that $1+G(\omega)=|U(\omega)|^2$, with $U(\omega)=u_0+\hat{\boldsymbol{u}}\varphi(\omega)$ and $\hat{\boldsymbol{u}}=[\,u_1\ u_2\ \ldots\ u_\nu\,]$. Then $\bar{I}(G(\omega))$ in (40) can be rewritten as

$$\bar{I}(G(\omega))=1+2\log(u_0)+\frac{1}{2\pi}\int_{-\pi}^{\pi}M(\omega)\big(u_0^2+2\mathrm{Re}\{u_0\hat{\boldsymbol{u}}\varphi(\omega)\}+\hat{\boldsymbol{u}}\varphi(\omega)\varphi^{\mathrm{H}}(\omega)\hat{\boldsymbol{u}}^{\mathrm{H}}\big)\mathrm{d}\omega. \tag{43}$$

Taking the first order differentials with respect to u_0 and $\hat{\boldsymbol{u}}$ and optimizing them directly results in the optimal solution (41). Inserting (41) back into (43) and after some manipulations, the optimal asymptotic rate is then in (42). ■

A. Method I

The structures of matrices $(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F})$ have the same constraints as in Section IV-A, except that now the matrices have infinite dimensions. Applying Szegö's theorem to (8), the asymptotic rate reads,

$$\begin{aligned}\bar{I}\big(W(\omega),T(\omega),F(\omega)\big) &= \lim_{K\to\infty}\frac{1}{K}I_{\mathrm{GMI}}(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F})\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\bigg(\log\big(1+|F(\omega)|^2\big)-|F(\omega)|^2-\frac{L_1(\omega)}{1+|F(\omega)|^2}\bigg)\mathrm{d}\omega\\ &\quad+\frac{1}{\pi}\int_{-\pi}^{\pi}\mathrm{Re}\big\{F^*(\omega)\big(W(\omega)H(\omega)-\alpha T(\omega)\big)\big\}\mathrm{d}\omega \qquad (44)\end{aligned}$$

where

$$L_1(\omega)=|F(\omega)W(\omega)|^2\big(N_0+|H(\omega)|^2\big)+\alpha|F(\omega)T(\omega)|^2-2\alpha|F(\omega)|^2\mathrm{Re}\{H(\omega)W(\omega)T^*(\omega)\}$$

and $H(\omega)$, $F(\omega)$, $W(\omega)$ and $T(\omega)$ are Fourier series associated to the band shaped Toeplitz matrices $\boldsymbol{H}$, $\boldsymbol{F}$, $\boldsymbol{W}$ and $\boldsymbol{T}$, respectively.

Applying Szegö's eigenvalue distribution theorem, the Fourier series associated to $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$ defined in (9) and (10) are

$$M(\omega) = \frac{|H(\omega)|^2}{N_0+|H(\omega)|^2}-1, \qquad (45)$$

$$\tilde{M}(\omega) = \alpha^2(M(\omega)+1)-\alpha. \qquad (46)$$

Define a $(2N_{\mathrm{T}}-\nu)\times 1$ vector

$$\phi(\omega)=\big[\exp\big(-jN_{\mathrm{T}}\omega\big)\ \ldots\ \exp\big(-j\omega\big)\ \exp\big(j(\nu+1)\omega\big)\ \ldots\ \exp\big(jN_{\mathrm{T}}\omega\big)\big]^{\mathrm{T}}, \qquad (47)$$

a $(2N_{\mathrm{T}}-\nu)\times 1$ vector $\boldsymbol{\varepsilon}_1$, and a $(2N_{\mathrm{T}}-\nu)\times(2N_{\mathrm{T}}-\nu)$ Hermitian matrix $\boldsymbol{\varepsilon}_2$ as

$$\begin{aligned}\boldsymbol{\varepsilon}_1 &= \frac{\alpha}{2\pi}\int_{-\pi}^{\pi}M(\omega)F^*(\omega)\phi(\omega)\mathrm{d}\omega,\\ \boldsymbol{\varepsilon}_2 &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|F(\omega)|^2\phi(\omega)\phi(\omega)^{\mathrm{H}}}{1+|F(\omega)|^2}\mathrm{d}\omega, \qquad (48)\end{aligned}$$

where N_{T} is the tap length of $T(\omega)$ and $\nu+1$ is the band size where matrix $\boldsymbol{T}$ is constrained to zero. Then we have Proposition 5 with the proof given in Appendix H.

Proposition 5. *The optimal $W(\omega)$ for the asymptotic rate in (44) is,*

$$W_{\mathrm{opt}}(\omega) = \frac{H^*(\omega)}{F^*(\omega)(N_0+|H(\omega)|^2)}\big(1+|F(\omega)|^2+\alpha F^*(\omega)T_{\mathrm{opt}}(\omega)\big), \qquad (49)$$

and when $0<\alpha\leq 1$, the optimal $T(\omega)$ in (44) is,

$$T_{\rm opt}(\omega) = -\boldsymbol{\varepsilon}_1^{\rm H}\boldsymbol{\varepsilon}_2^{-1}\phi(\omega). \tag{50}$$

With the optimal $W(\omega)$ and $T(\omega)$, the asymptotic rate reads,

$$\bar{I}\big(W_{\rm opt}(\omega), T_{\rm opt}(\omega), F(\omega)\big) = \begin{cases} \bar{I}_1(F(\omega)), & \alpha=0 \\ \bar{I}_1(F(\omega)) + \bar{\delta}_1(F(\omega)), & 0<\alpha\leq 1. \end{cases} \tag{51}$$

The functions $\bar{I}_1(F(\omega))$ and $\bar{\delta}_1(F(\omega))^5$ are defined as,

$$\bar{I}_1(F(\omega)) = 1 + \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+|F(\omega)|^2\big)+M(\omega)\big(1+|F(\omega)|^2\big)\Big){\rm d}\omega, \tag{52}$$

$$\bar{\delta}_1(F(\omega)) = -\boldsymbol{\varepsilon}_1^{\rm H}\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\varepsilon}_1. \tag{53}$$

A closed form solution of the optimal $F(\omega)$ in (51) seems out of reach and a gradient based optimization is therefore used. Note that if we replace $|F(\omega)|^2$ by $|G(\omega)|$, (52) has the same form as (40) in Theorem 3, and the optimal solution of $G(\omega)$ is in closed form. However, as the optimal $G(\omega)$ can not always be decomposed as $G(\omega)=|F(\omega)|^2$, (52) also needs a numerical optimization whenever $G(\omega)$ is not positive real values for all ω. The starting point initialization for the optimization procedure is similar as in the finite linear vector channels. In the ISI case, Method I is still not concave, an example is also provided in Appendix C.

We next derive the differentials of the optimal asymptotic rate with respect to $F(\omega)$ in (51). With memory constraint ν, the Fourier series associated to $\boldsymbol{F}$ is

$$F(\omega) = \sum_{k=0}^{\nu} f_k \exp\big(jk\omega\big)$$

where ν is the memory length and f_k is the non-zero element at the kth lower diagonal of $\boldsymbol{F}$. The differential of $\bar{I}_1(F(\omega))$ with respect to f_k is

$$\frac{\partial \bar{I}_1(F(\omega))}{\partial f_k} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\bigg(M(\omega)+\frac{1}{1+|F(\omega)|^2}\bigg)\sum_{m=0}^{\nu} f_m^* \exp\big(j(k-m)\omega\big){\rm d}\omega,$$

and the differential of $\bar{\delta}_1(F(\omega))$ with respect to f_k is

$$\frac{\partial \bar{\delta}_1(F(\omega))}{\partial f_k} = -\frac{\partial \boldsymbol{\varepsilon}_1^{\rm H}}{\partial f_k}\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_1^{\rm H}\boldsymbol{\varepsilon}_2^{-1}\frac{\partial \boldsymbol{\varepsilon}_2}{\partial f_k}\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\varepsilon}_1,$$

[5] Similar to finite length linear vector channels, $\bar{\delta}_1(F(\omega))$ in (53) is only defined for $\alpha\neq 0$, as when $\alpha=0$, $\tilde{M}(\omega)=0$ and the inversion part in $\bar{\delta}_1(F(\omega))$ is not well defined. The same comment holds for $\bar{\delta}_2(G(\omega))$ in (63).

where

$$\frac{\partial \boldsymbol{\varepsilon}_1^{\mathrm{H}}}{\partial f_k} = \frac{\alpha}{2\pi}\int_{-\pi}^{\pi} M(\omega)\phi(\omega)^{\mathrm{H}}\exp(jk\omega)\mathrm{d}\omega$$

$$\frac{\partial \boldsymbol{\varepsilon}_2}{\partial f_k} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\tilde{M}(\omega)\phi(\omega)\phi(\omega)^{\mathrm{H}}}{\big(1+|F(\omega)|^2\big)^2}\sum_{m=0}^{\nu} f_m^{*}\exp\big(j(k-m)\omega\big)\mathrm{d}\omega.$$

The connection between the optimal front-end filter and the interference cancelation matrix as in Proposition 2 also holds for ISI channels. The asymptotic version of Proposition 2 that shows the relationship between the optimal $W(\omega)$ and $T(\omega)$ is stated in Proposition 6.

Proposition 6. *When $0<\alpha\leq 1$, define the Fourier transforms*

$$a_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} F^{*}(\omega)W_{\mathrm{opt}}(\omega)H(\omega)\exp\big(-jk\omega\big)\mathrm{d}\omega$$

$$b_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} F^{*}(\omega)T_{\mathrm{opt}}(\omega)\exp\big(-jk\omega\big)\mathrm{d}\omega,$$

then $a_k=b_k$ holds for $k<-(\nu+1)$.

Proof: In Appendix H, the optimal $\tilde{\boldsymbol{t}}$ in (94) satisfies

$$\tilde{\boldsymbol{t}}_{\mathrm{opt}}\boldsymbol{\varepsilon}_2 = -\boldsymbol{\varepsilon}_1^{\mathrm{H}}.$$

With the definitions of $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$ in (48), this is equivalent to

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|F(\omega)|^2 T_{\mathrm{opt}}(\omega)\phi(\omega)^{\mathrm{H}}}{1+|F(\omega)|^2}\mathrm{d}\omega = -\frac{\alpha}{2\pi}\int_{-\pi}^{\pi}F(\omega)M(\omega)\phi(\omega)^{\mathrm{H}}\mathrm{d}\omega. \tag{54}$$

On the other hand, with W_{opt} in (49) and $M(\omega)$, $\tilde{M}(\omega)$ defined in (45) and (46), we have

$$\begin{aligned}&\frac{1}{2\pi}\int_{-\pi}^{\pi}(F^{*}(\omega)W_{\mathrm{opt}}(\omega)H(\omega)-F^{*}(\omega)T_{\mathrm{opt}}-\big(1+|F(\omega)|^2\big))\exp\big(-jk\omega\big)\mathrm{d}\omega\\ &\qquad= \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\frac{\tilde{M}(\omega)F^{*}(\omega)T_{\mathrm{opt}}(\omega)}{\alpha}+\big(1+|F(\omega)|^2\big)M(\omega)\Big)\exp\big(-jk\omega\big)\mathrm{d}\omega. \end{aligned}\tag{55}$$

Transforming (54) and (55) back into matrix forms, we have that (19) and (20) hold. Following the same arguments as in the proof of Proposition 2, $\boldsymbol{F}^{\mathrm{H}}(\boldsymbol{W}_{\mathrm{opt}}\boldsymbol{H}-\boldsymbol{R}_{\mathrm{opt}})$ is banded within diagonals $[-\nu, K-1]$. Therefore we have

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\big(F^{*}(\omega)W_{\mathrm{opt}}(\omega)H(\omega)-F^{*}(\omega)T_{\mathrm{opt}}(\omega)\big)\exp\big(-jk\omega\big)\mathrm{d}\omega=0$$

whenever $k<-(\nu+1)$, which proves Proposition 6. ■

B. Method II

The structures of matrices $(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G})$ in ISI case have the same constraints as in Section IV-B while the dimensions of these matrices are infinitely large. The type (a) $\boldsymbol{R}$ in Figure 3 is not meaningful as $N,K\to\infty$. We use the same definition of ν_{R} as in Proposition 3. That is, ν_{R} equals 0 or ν, corresponding to type (b) or type (c) of $\boldsymbol{R}$ as shown in Figure 3. Type (a) is not considered for ISI.

Applying Szegö's theorem to (6), the asymptotic rate for Method II is

$$\begin{aligned}\bar{I}(V(\omega),R(\omega),G(\omega)) &= \lim_{K\to\infty}\frac{1}{K}I_{\mathrm{GMI}}(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G})\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\log\big(1+G(\omega)\big)-G(\omega)-\frac{L_2(\omega)}{1+G(\omega)}\right)\mathrm{d}\omega\\ &\quad+\frac{1}{\pi}\int_{-\pi}^{\pi}\mathrm{Re}\left\{\big(V(\omega)H(\omega)-\alpha R(\omega)\big)\right\}\mathrm{d}\omega \end{aligned}\tag{56}$$

where

$$L_2(\omega)=|V(\omega)|^2\big(N_0+|H(\omega)|^2\big)+\alpha|R(\omega)|^2-2\alpha\mathrm{Re}\big\{H(\omega)V(\omega)R^*(\omega)\big\}$$

and $V(\omega)$, $R(\omega)$ and $G(\omega)$ are Fourier series associated to the band shaped Toeplitz matrices $\boldsymbol{V}$, $\boldsymbol{R}$, and $\boldsymbol{G}$, respectively.

Define a $2(N_{\mathrm{R}}-\nu_{\mathrm{R}})\times 1$ vector

$$\psi(\omega)=\big[\exp\!\big(-jN_{\mathrm{R}}\omega\big)\ \ldots\ \exp\!\big(-j(\nu_{\mathrm{R}}+1)\omega\big)\ \exp\!\big(j(\nu_{\mathrm{R}}+1)\omega\big)\ \ldots\ \exp\!\big(jN_{\mathrm{R}}\omega\big)\big]^{\mathrm{T}},\tag{57}$$

a $2(N_{\mathrm{R}}-\nu_{\mathrm{R}})\times 1$ vector $\boldsymbol{\zeta}_1$, and a $2(N_{\mathrm{R}}-\nu_{\mathrm{R}})\times 2(N_{\mathrm{R}}-\nu_{\mathrm{R}})$ Hermitian matrix $\boldsymbol{\zeta}_2$ as

$$\begin{aligned}\boldsymbol{\zeta}_1 &= \frac{\alpha}{2\pi}\int_{-\pi}^{\pi}M(\omega)\psi(\omega)\mathrm{d}\omega,\\ \boldsymbol{\zeta}_2 &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)\psi(\omega)\psi(\omega)^{\mathrm{H}}}{1+G(\omega)}\mathrm{d}\omega,\end{aligned}\tag{58}$$

where N_{R} denotes the tap length of $R_{\mathrm{opt}}(\omega)$, $2\nu_{\mathrm{R}}+1$ is the band size where $\boldsymbol{R}$ is constrained to zero, and $M(\omega)$ and $\tilde{M}(\omega)$ are defined in (45) and (46). Then with this notation, we have

Proposition 7. *The optimal $V(\omega)$ for (56) is,*

$$V_{\mathrm{opt}}(\omega)=\frac{H^*(\omega)}{N_0+|H(\omega)|^2}\big(1+G(\omega)+\alpha R_{\mathrm{opt}}(\omega)\big),\tag{59}$$

and when $0<\alpha\leq 1$, the optimal $R(\omega)$ for the asymptotic rate in (56) is,

$$R_{\mathrm{opt}}(\omega)=-\boldsymbol{\zeta}_1^{\mathrm{H}}\boldsymbol{\zeta}_2^{-1}\psi(\omega).\tag{60}$$

With the optimal $V(\omega)$ and $R(\omega)$, the asymptotic rate reads,

$$\bar{I}\big(V_{\text{opt}}(\omega), R_{\text{opt}}(\omega), G(\omega)\big) = \begin{cases} \bar{I}_2(G(\omega)), & \alpha = 0 \\ \bar{I}_2(G(\omega)) + \bar{\delta}_2(G(\omega)), & 0 < \alpha \le 1. \end{cases} \tag{61}$$

The functions $\bar{I}_1(G(\omega))$ and $\bar{\delta}_2(G(\omega))$ are defined as,

$$\bar{I}_2(G(\omega)) = 1 + \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+G(\omega)\big) + M(\omega)\big(1+G(\omega)\big)\Big)\mathrm{d}\omega, \tag{62}$$

$$\bar{\delta}_2(G(\omega)) = -\boldsymbol{\zeta}_1^{\mathrm{H}}\boldsymbol{\zeta}_2^{-1}\boldsymbol{\zeta}_1. \tag{63}$$

The proof is given in Appendix I, where we also show that $R(\omega)$ is real and the matrix $\boldsymbol{R}$ has Hermitian symmetry.

The optimal $G(\omega)$ for (62) can be solved in closed form from Theorem 3. However, a closed form solution for the optimal $G(\omega)$ in (61) when $0 < \alpha \le 1$ again seems out of reach and a gradient based optimization is used. The asymptotic rate $\bar{I}(V_{\text{opt}}(\omega), R_{\text{opt}}(\omega), G(\omega))$ in Method II is also concave with respect to $G(\omega)$, the proof is provided in Appendix J.

Below we derive the differentials of the optimal asymptotic rate with respect to $G(\omega)$ for $0<\alpha\le 1$. As matrix $\boldsymbol{G}$ is Hermitian, the associated Fourier series $G(\omega)$ is

$$G(\omega) = \sum_{k=-\nu}^{\nu} g_k \exp(jk\omega) = g_0 + 2\mathrm{Re}\Big\{\sum_{k=1}^{\nu} g_k \exp\big(jk\omega\big)\Big\} \tag{64}$$

where g_k is the element at the kth diagonal of $\boldsymbol{G}$. The differential of $\bar{I}\big(V_{\text{opt}}(\omega), R_{\text{opt}}(\omega), G(\omega)\big)$ with respect to g_k is

$$\frac{\partial \bar{I}}{\partial g_k} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(M(\omega) + \frac{1}{1+G(\omega)}\Big)\exp(jk\omega)\mathrm{d}\omega + \boldsymbol{\zeta}_1^{\mathrm{H}}\boldsymbol{\zeta}_2^{-1}\frac{\partial \boldsymbol{\zeta}_2}{\partial g_k}\boldsymbol{\zeta}_2^{-1}\boldsymbol{\zeta}_1$$

where

$$\frac{\partial \boldsymbol{\zeta}_2}{\partial g_k} = -\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)\psi(\omega)\psi(\omega)^{\mathrm{H}}}{\big(1+G(\omega)\big)^2}\exp\big(jk\omega\big)\mathrm{d}\omega.$$

The asymptotic version of Proposition 4 that shows the relationship between the optimal $V(\omega)$ and $R(\omega)$ for ISI channels is stated in Proposition 8.

Proposition 8. *When $0<\alpha\le 1$, define the Fourier transforms*

$$a_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} V_{\text{opt}}(\omega)H(\omega)\exp\big(-jk\omega\big)\mathrm{d}\omega, \tag{65}$$

$$b_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} R_{\text{opt}}(\omega)\exp\big(-jk\omega\big)\mathrm{d}\omega, \tag{66}$$

then $a_k = b_k$ holds for $|k| > \nu + \nu_{\rm R}$, where $\nu_{\rm R} = 0$ for Method II.b and $\nu_{\rm R} = \nu$ for Method II.c.

Proof: In Appendix I, the optimal $\tilde{\boldsymbol{r}}$ given in (99) satisfies,

$$\tilde{\boldsymbol{r}}_{\rm opt}\boldsymbol{\zeta}_2 = -\boldsymbol{\zeta}_1^{\rm H}. \tag{67}$$

With the definitions of in $\boldsymbol{\zeta}_1$, $\boldsymbol{\zeta}_2$ in (58), (67) is equivalent to

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)R_{\rm opt}(\omega)\psi(\omega)^{\rm H}}{1+G(\omega)}{\rm d}\omega = -\frac{\alpha}{2\pi}\int_{-\pi}^{\pi}M(\omega)\psi(\omega)^{\rm H}{\rm d}\omega. \tag{68}$$

On the other hand, with $V_{\rm opt}(\omega)$ in (59) and $M(\omega)$, $\tilde{M}(\omega)$ defined in (45) and (46), we have

$$\begin{aligned}&\frac{1}{2\pi}\int_{-\pi}^{\pi}\big(V_{\rm opt}(\omega)H(\omega) - R_{\rm opt}(\omega) - (1+G(\omega))\big)\exp\big(-jk\omega\big){\rm d}\omega\\ &\quad = \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\frac{\tilde{M}(\omega)R_{\rm opt}(\omega)}{\alpha} + M(\omega)\big(1+G(\omega)\big)\Big)\exp\big(-jk\omega\big){\rm d}\omega.\end{aligned} \tag{69}$$

Transforming (68) and (69) back into matrix forms, we have that (27) and (28) hold. Following the same arguments as in the proof of Proposition 4, $\boldsymbol{V}_{\rm opt}\boldsymbol{H} - \boldsymbol{R}_{\rm opt}$ is banded within diagonals $[-(\nu+\nu_{\rm R}), \nu+\nu_{\rm R}]$. Therefore we have

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\big(V_{\rm opt}(\omega)H(\omega) - R_{\rm opt}(\omega)\big)\exp\big(-jk\omega\big){\rm d}\omega = 0$$

whenever $|k| > \nu + \nu_{\rm R} + 1$, which proves Proposition 8. ■

We provide an example to illustrate Proposition 8 in Figure 5 with Method II.c and $\nu = \nu_{\rm R} = 1$. The Proakis-C [43] channel is tested at an SNR of 10 dB and α, which represents the soft information feedback quality, equals 0.1, 0.4 and 0.8, respectively. Since $\nu_{\rm R} = 1$, b_k as defined in (66) is constrained to zero for $0 \leq k \leq 1$. As can be seen, a_k as defined in (65) equals b_k only for $|k| > 2$, and when $|k| = 2$, a_k and b_k are not identical. This shows that with the optimal $V(\omega)$ and $R(\omega)$, the signal part along the second upper and lower diagonals that is not considered in $G(\omega)$ shall not be perfectly canceled out. This behavior cannot be seen in [44], which treats LMMSE-PIC for ISI channels, since $\nu = \nu_{\rm R} = 0$.

C. Method III

Similar as finite length linear vector channels, we also investigate Method III for ISI channels. Applying Szegö's theorem to (38), the asymptotic rate reads,

$$\begin{aligned}\bar{I}(G(\omega)) &= \lim_{K\to\infty}\frac{1}{K}I_{\rm GMI}(\boldsymbol{G})\\ &= 1 + \frac{1}{2\pi}\int_{-\pi}^{\pi}\big(\log\big(1+G(\omega)\big) + \hat{M}(\omega)(1+G(\omega))\big){\rm d}\omega\end{aligned} \tag{70}$$

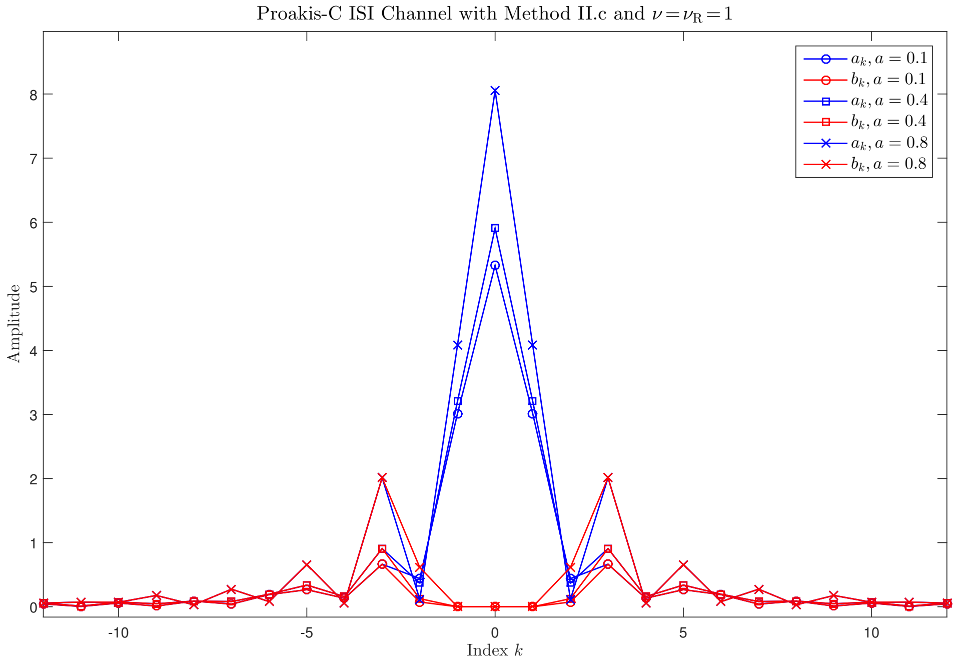

Fig. 5: Comparison between a_k and b_k for Proakis-C channel with Method II.c.

where

$$\hat{M}(\omega) = 2\mathrm{Re}\left\{\alpha \hat{W}(\omega) H(\omega) \hat{C}^*(\omega) + \hat{W}(\omega) H(\omega) - \alpha \hat{C}^*(\omega)\right\} - \frac{|\hat{W}(\omega)|^2}{N_0 + |H(\omega)|^2} - \alpha |\hat{C}(\omega)|^2 - 1 \quad (71)$$

and $\hat{W}(\omega)$, $\hat{C}(\omega)$ are Fourier series associated to the band shaped Toeplitz matrices $\hat{\boldsymbol{W}}$, $\hat{\boldsymbol{C}}$ defined in (35) and (36), respectively. As (70) has the same form as (40) in Theorem 3, replacing $M(\omega)$ by $\hat{M}(\omega)$ in (40), the optimal $G(\omega)$ and asymptotic rate follows directly from Theorem 3.

Remark 4. *Proposition 8 also holds for Method III with $\nu_{\mathrm{R}} = \nu$. This comes from the fact that $[(\boldsymbol{I} + \boldsymbol{G})(\hat{\boldsymbol{W}}\boldsymbol{H} - \hat{\boldsymbol{C}})]_{\backslash 2\nu} = \boldsymbol{0}$.*

VI. SNR Asymptotics

In this section, we analyze asymptotic properties of the CS demodulators as N_0 goes to 0 and ∞. As Method I is inferior to Method II in GMI sense, we limit our investigations to Method II and Method III. We start the analysis for finite length linear vector channels first and with the following limits that can be verified straightforwardly:

$$\lim_{N_0\to 0} \boldsymbol{M}/N_0 = -(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1},$$
$$\lim_{N_0\to\infty} N_0(\boldsymbol{I}+\boldsymbol{M}) = \boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}. \tag{72}$$

Moreover, we also have

$$\lim_{N_0\to 0} \tilde{\boldsymbol{M}} = \boldsymbol{P}^2 - \boldsymbol{P},$$
$$\lim_{N_0\to\infty} \tilde{\boldsymbol{M}} = -\boldsymbol{P}. \tag{73}$$

Note that when $\boldsymbol{P}\neq\boldsymbol{0}$, it must hold by the definition of $\delta_2(\boldsymbol{G})$ in (25), that $\tilde{\boldsymbol{M}}$ is invertible. As $N_0 \to 0$, $\tilde{\boldsymbol{M}}\to\boldsymbol{P}^2-\boldsymbol{P}$, which implies that $\boldsymbol{P}\prec\boldsymbol{I}$. Therefore, we restrict the asymptotic SNR analysis to $\boldsymbol{P}\prec\boldsymbol{I}$.

Lemma 3. *When $N_0\to 0$ and ∞, the optimal $\boldsymbol{G}$ for (23) in Method II satisfies (17), and the following limits hold,*

$$\lim_{N_0\to 0} \left[(N_0(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}}))^{-1}\right]_\nu = [(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1}]_\nu, \tag{74}$$
$$\lim_{N_0\to\infty} [N_0\boldsymbol{G}_{\mathrm{opt}}]_\nu = [\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}]_\nu. \tag{75}$$

Proof: When $\boldsymbol{P}=\boldsymbol{0}$, from Theorem 2 the optimal $\boldsymbol{G}$ for (23) satisfies (17). Next we prove that, when $\boldsymbol{P}\neq\boldsymbol{0}$, as $N_0\to 0$ and $N_0\to\infty$, the gradient of $\delta_2(\boldsymbol{G})$ in (25) converges to zero, therefore (17) also holds. From (72), when $N_0\to 0$, $\boldsymbol{M}\to\boldsymbol{0}$ and $N_0\to\infty$, $\boldsymbol{M}\to-\boldsymbol{I}$. Therefore, by the definition of $\boldsymbol{\Omega}$,

$$\lim_{N_0\to 0,\infty} \boldsymbol{d} = \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{M}\boldsymbol{P}) = \boldsymbol{0}.$$

This implies that the gradient $d_{\boldsymbol{G}}(\delta_2)$ in (89) (see Appendix F) converges to zero. Hence the differentials of $I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R}_{\mathrm{opt}},\boldsymbol{G})$ in (23) when $\boldsymbol{P}\neq\boldsymbol{0}$ converges to the differentials when $\boldsymbol{P}=\boldsymbol{0}$, which proves the first part of the lemma.

From (17) and (72), the limit (74) follows and

$$\lim_{N_0\to\infty} \left[N_0\left(\boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})^{-1}\right)\right]_\nu = \lim_{N_0\to 0}[N_0(\boldsymbol{I}+\boldsymbol{M})]_\nu = [\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}]_\nu. \tag{76}$$

Therefore, $\boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})^{-1}\to\boldsymbol{0}$[6] as $N_0\to\infty$. By the matrix inversion lemma, $\boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})^{-1}\to\boldsymbol{G}_{\mathrm{opt}}$ as $N_0\to\infty$, and combining this with (76) proves the limit (75). ■

[6] A matrix $\boldsymbol{A}\to\boldsymbol{B}$ or a vector $\boldsymbol{a}\to\boldsymbol{b}$ means the non-zero elements of $\boldsymbol{A}-\boldsymbol{B}$ or $\boldsymbol{a}-\boldsymbol{b}$ converges to zero.

Lemma 4. *In Method II, with the optimal $\boldsymbol{G}$, when $N_0 \to 0$ the GMI increment $\delta_2(\boldsymbol{G})$ in (25) converges to zero with speed $\mathcal{O}(1/N_0)$*[7] *and when $N_0 \to \infty$ the GMI increment $\delta_2(\boldsymbol{G})$ converges to zero with speed $\mathcal{O}(N_0^2)$.*

Proof: As $N_0 \to 0$, from (72) we have

$$\lim_{N_0\to 0} \boldsymbol{d}/N_0 = \lim_{N_0\to 0} \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{M}\boldsymbol{P}/N_0) = -\boldsymbol{\Omega}\mathrm{vec}\big((\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1}\boldsymbol{P}\big).$$

Based on (73) and Lemma 3, the below equalities holds,

$$\delta_2(\boldsymbol{G}_{\mathrm{opt}}) = N_0 \frac{\boldsymbol{d}^{\mathrm{H}}}{N_0}\Big(\boldsymbol{\Omega}\Big(\tilde{\boldsymbol{M}}^{*} \otimes \frac{(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})^{-1}}{N_0}\Big)\boldsymbol{\Omega}^{\mathrm{T}}\Big)^{-1}\frac{\boldsymbol{d}}{N_0} = \mathcal{O}(N_0).$$

On the other hand, as $N_0 \to \infty$, by the definition of $\boldsymbol{\Omega}$, from (72) we have

$$\lim_{N_0\to\infty} N_0\boldsymbol{d} = \lim_{N_0\to\infty} \boldsymbol{\Omega}\mathrm{vec}(N_0\boldsymbol{M}\boldsymbol{P}) = \lim_{N_0\to\infty} \boldsymbol{\Omega}\mathrm{vec}\big(N_0(\boldsymbol{I}+\boldsymbol{M})\boldsymbol{P}\big) = \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}\boldsymbol{P}).$$

Based on (73) and Lemma 3, the below equalities holds,

$$\delta_2(\boldsymbol{G}_{\mathrm{opt}}) = \frac{1}{N_0^2}(N_0\boldsymbol{d}^{\mathrm{H}})\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*} \otimes (\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}\big)^{-1}(N_0\boldsymbol{d}) = \mathcal{O}(1/N_0^2).$$

Therefore, Lemma 4 holds. ■

Lemma 5. *When $N_0 \to 0$ and ∞, the optimal GMI in Method III is independent of $\boldsymbol{P}$ and converges to the optimal GMI for $\boldsymbol{P} = \boldsymbol{0}$. Moreover, (74) and (75) hold.*

The proof is given in Appendix K. Combining Lemmas 3-5 and noticing the fact that Method III and Method II are equivalent when $\boldsymbol{P} = \boldsymbol{0}$, we have the following Theorem 4.

Theorem 4. *Assume that $\boldsymbol{P} \prec \boldsymbol{I}$, when $N_0 \to 0$ and ∞, the optimal GMI in Method III converges to the optimal GMI in Method III with $\boldsymbol{P} = \boldsymbol{0}$. Moreover, the optimal GMI in Method II also converges to the optimal GMI in Method III with $\boldsymbol{P} = \boldsymbol{0}$, with speed $\mathcal{O}(1/N_0)$ when SNR increase and $\mathcal{O}(N_0^2)$ when SNR decreases. The optimal $\boldsymbol{G}$ for both methods has the following asymptotic properties:*

$$\lim_{N_0\to 0}[(N_0(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}}))^{-1}]_{\nu} = [(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1}]_{\nu},$$

[7] Two scalars A and B as functions of a variable n converging to each other with speed $\mathcal{O}(n)$ means that, there exists a constant C such that $\lim_{n\to\infty} n|A-B| < C$.

$$\lim_{N_0\to\infty}[N_0\boldsymbol{G}_{\rm opt}]_\nu = [\boldsymbol{H}^{\rm H}\boldsymbol{H}]_\nu.$$

From Theorem 4 we know that, except for the case where one of the elements in the diagonal matrix $\boldsymbol{P}$ is 1, the soft feedback information becomes asymptotically insignificant for the design of the CS parameters. The reason is that, when $N_0 \to 0$, $\hat{\boldsymbol{x}}$ is overwhelmed by the noise, while when $N_0 \to \infty$, the optimal front-end filter will null out $\hat{\boldsymbol{x}}$ since the filter can perfectly reconstruct the transmitted symbols without using the side information.

Remark 5. *When $N_0 \to 0$, the optimal CS demodulator is the EZF demodulator as defined in Example 1, and when $N_0 \to \infty$, the optimal CS demodulator becomes the TMF as defined in Example 2.*

Next we extend Theorem 4 to ISI channels. With ISI channels, as the same constraint $\boldsymbol{P} = \alpha\boldsymbol{I} \prec \boldsymbol{I}$ shall hold, we make the restriction that $0 \le \alpha < 1$. Taking the trace on both sides of the equations in (72) and (73), and Applying Szegö's eigenvalue distribution theorem, we obtain the following limits:

$$\begin{aligned}
&\lim_{N_0\to 0}\int_{-\pi}^{\pi}\frac{M(\omega)}{N_0}{\rm d}\omega = -\int_{-\pi}^{\pi}\frac{1}{|H(\omega)|^2}{\rm d}\omega,\\
&\lim_{N_0\to\infty}\int_{-\pi}^{\pi}N_0(1+M(\omega)){\rm d}\omega = \int_{-\pi}^{\pi}|H(\omega)|^2{\rm d}\omega,\\
&\lim_{N_0\to 0}\int_{-\pi}^{\pi}\tilde{M}(\omega){\rm d}\omega = \alpha(\alpha-1),\\
&\lim_{N_0\to\infty}\int_{-\pi}^{\pi}\tilde{M}(\omega){\rm d}\omega = -\alpha.
\end{aligned}\tag{77}$$

With the above limits in (77), the SNR asymptotic properties for ISI channels are presented in Corollary 1, which is an asymptotic version of Theorem 4 when the channel matrix $\boldsymbol{H}$ and CS parameters are band shaped Toeplitz matrices with infinite dimensions. The detailed proof is following the same analysis as for the finite linear vector channels and omitted.

Corollary 1. *Assume that $0 \le \alpha < 1$, when $N_0 \to 0$ and ∞, the optimal GMI in Method III converges to the optimal GMI in Method III with $\alpha = 0$. Moreover, the optimal GMI in Method II also converges to the optimal GMI in Method III with $\alpha = 0$, with speed $\mathcal{O}(1/N_0)$ when SNR increase and $\mathcal{O}(N_0^2)$ when SNR decreases. The optimal $\boldsymbol{G}$ for both methods has the following* *asymptotic properties hold for $|k|\leq\nu$:*

$$\lim_{N_0\to 0}\int_{-\pi}^{\pi}\frac{1}{N_0(1+G_{\mathrm{opt}}(\omega))}\exp(-jk\omega)\mathrm{d}\omega=\int_{-\pi}^{\pi}\frac{1}{|H(\omega)|^2}\exp(-jk\omega)\mathrm{d}\omega,$$

$$\lim_{N_0\to\infty}\int_{-\pi}^{\pi}N_0G_{\mathrm{opt}}(\omega)\exp(-jk\omega)\mathrm{d}\omega=\int_{-\pi}^{\pi}|H(\omega)|^2\exp(-jk\omega)\mathrm{d}\omega.$$

VII. Numerical Results

A. GMI Simulation for MIMO and ISI Channels

We first evaluate the GMIs for all CS demodulators with various feedback quality and with memory constraint $\nu=1$. In the 5×5 MIMO case, all channel elements are assumed to be independent identically distributed (IID) complex Gaussian and the received signal power at each receive antenna is normalized to unity. For the ISI case, we consider IID complex Gaussian channels with tap length $L=5$ and the total average power is normalized to unity. We simulated 10^4 channel realizations for each signal to noise ratio (SNR). The GMIs are compared with the GMI of the static CS demodulator used in [18] (denoted as "StaticCS" in the figures) which is equivalent to our case when $\boldsymbol{P}=\boldsymbol{0}$. The channel capacity[8] is also presented for reference.

The simulation results are plotted in Figure 6 and Figure 7. When the quality of the soft information improves beyond $\boldsymbol{P}=\boldsymbol{0}$, Method II.b performs the best among all CS demodulators since it has the most DoFs. Method II.c is the worst among Method I and Method II CS demodulators. Method I is slightly worse than Method II.a, which is because although the interference cancelation matrix $\boldsymbol{R}$ is type (a) in both cases, $\boldsymbol{R}$ in Method II.a is more general than in Method I since in Method I $\boldsymbol{R}$ is constrained to $\boldsymbol{R}=\boldsymbol{F}^{\mathrm{H}}\boldsymbol{T}$. The GMIs of Method III are inferior to Method II which is expected. However in the ISI case, it slightly outperforms Method II.c, which is because $\boldsymbol{R}$ in Method III has more doFs than Method II.c. The simulation results show consistent GMI increments for all CS demodulators when the feedback quality improves. When $\boldsymbol{P}$ increases from $\boldsymbol{P}=\boldsymbol{0}$ to the ideal case $\boldsymbol{P}=\boldsymbol{I}$, the channel capacity becomes inferior to the GMIs as the pair $(\hat{\boldsymbol{y}},\boldsymbol{x})$ becomes superior to $(\boldsymbol{y},\boldsymbol{x})$ for information transfer.

We also evaluate the SNR asymptotic properties described in Theorem 4 with IID complex Gaussian 5×5 MIMO channels. As showed in Figure 8, the GMIs of Method II.c and Method

[8] The channel capacity is calculated without any channel state information (CSI) at the TX side, therefore it does not contain any water-filling.

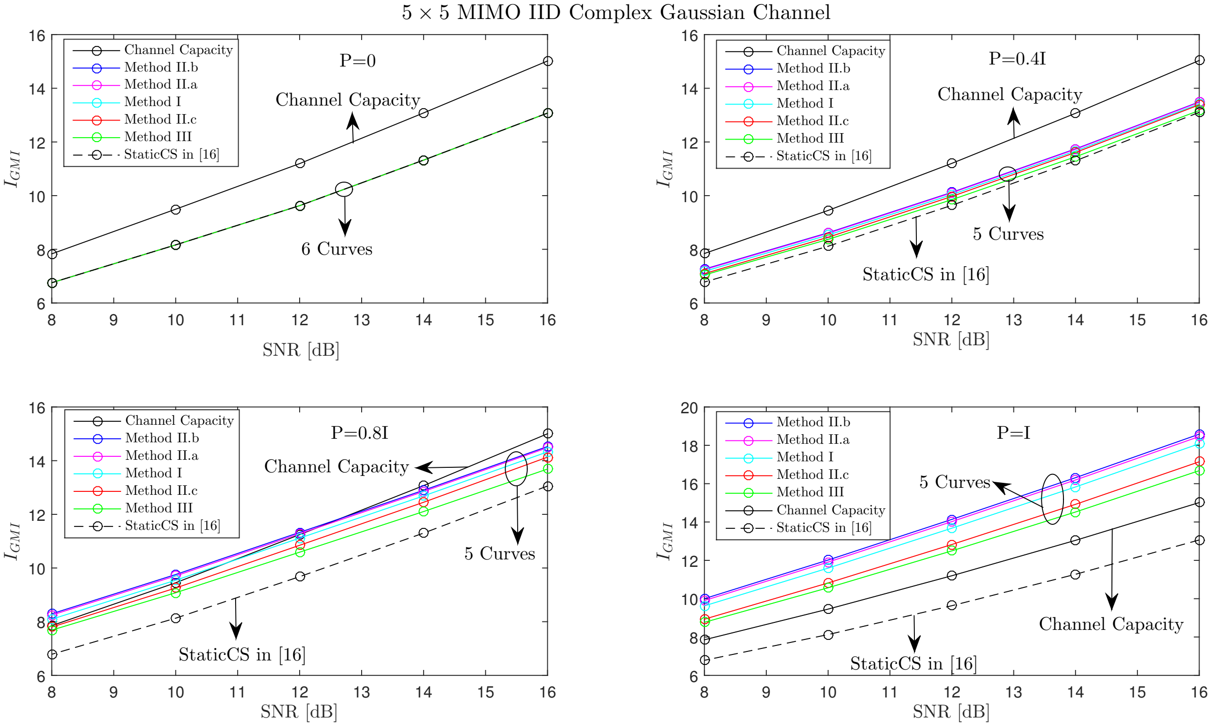

Fig. 6: GMI simulation results for 5×5 MIMO IID complex Gaussian channels with $\nu=1$.

III both converge to Method III with $\boldsymbol{P}=\mathbf{0}$. Moreover, the GMIs of CS demodulators converge to EZF at high SNR and TMF at low SNR, which is well aligned with Theorem 4.

Furthermore, in order to investigate the impact of permutation of the channel matrix, we simulate the GMI with permutations for 5×5 MIMO with Method II.c and $\nu=1$. The test set up is the same as in Figure 6. In order to obtain the optimal permutation, we exhaust all $5!=120$ possible permutations of the channel matrix for each channel realization, then we average the highest GMIs over all realizations. We also investigate an energy based method to select the optimal permutation. The energy based permutation is to choose the permutation over all 60 possible permutations (due to the Hermitian property of the Gram matrix $\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}$, the search space is reduced) that maximizes the total energy of the $2\nu+1$ diagonals in the Gram matrix where $\boldsymbol{G}$ can be non-zero. In Figure 9, it can be seen that the energy based permutation of the channels has reduced the gap of the GMIs between the non-permuted channels and the channels with optimal permutation.

We next turn to link-level simulations with a Turbo code [45] where the decoder uses 8 internal iterations. A single code block over all transmit symbols is used. The optimal scaling

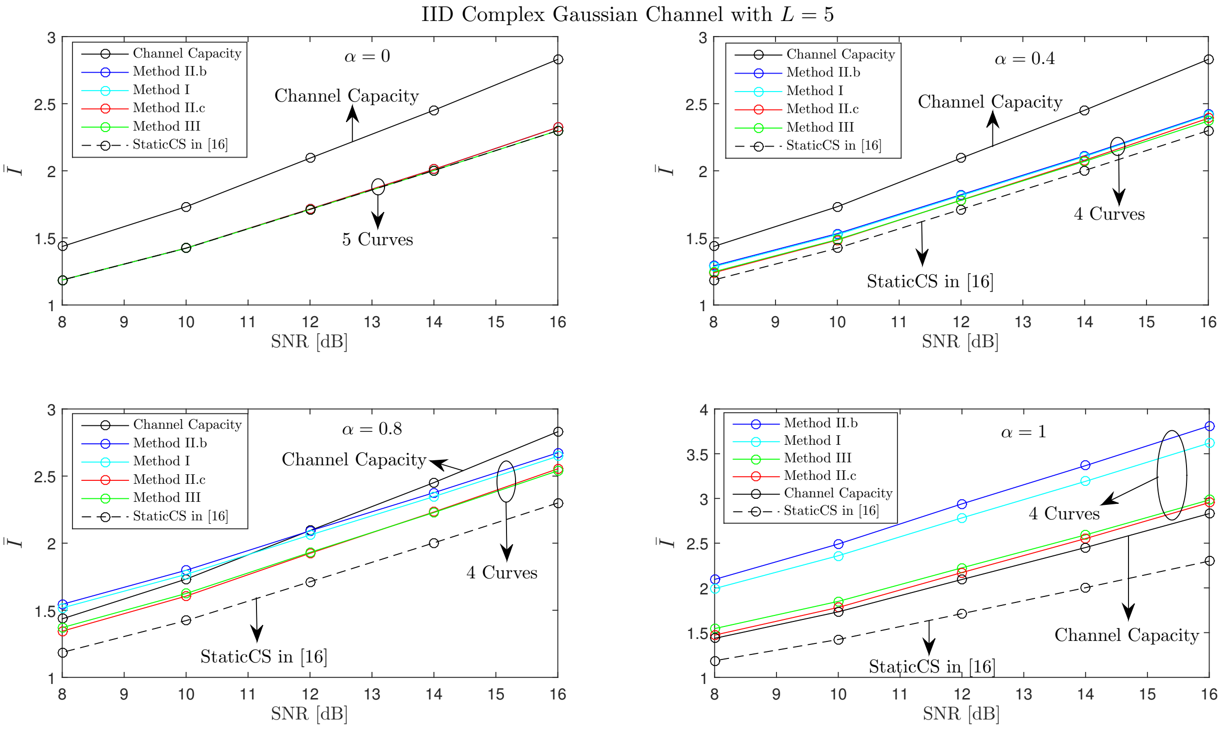

Fig. 7: GMI simulation results for IID complex Gaussian ISI channels with $\nu = 1$.

factor of the input and output extrinsic information is found by experiments to be 0.6 for the MAP demodulator. We use the same scaling factor for the CS demodulators. At each SNR point 10^4 coded blocks are simulated, while three global iterations are used between the demodulator and the decoder.

B. ISI Channels with Turbo Code

As an ISI channel a more natural setting for the CS demodulators, we start with simulations for ISI channels. The block error ratio (BLER) performance is used as metric for evaluating performance. The transmitted symbols are QPSK and 16QAM symbols. We choose $N_{\mathrm{T}} = N_{\mathrm{R}} = 8L$ for $T(\omega)$ and $R(\omega)$ in Method I and Method II, respectively, where L is the channel length. The number of the front-end filter taps in $W(\omega)$ and $V(\omega)$ are also set to $8L$.

In Figure 10 we plot the performance for Proakis-B channel [43] with QPSK symbols and a (1064, 1600) turbo code. The three taps are $\boldsymbol{h} = [\,0.407\; 0.815\; 0.407\,]$. As can be seen, the gap between the LMMSE-PIC and MAP demodulators is around 3-4 dB for all three iterations. With the proposed CS demodulators, the gap is reduced to less than 0.5 dB. Moreover, Method

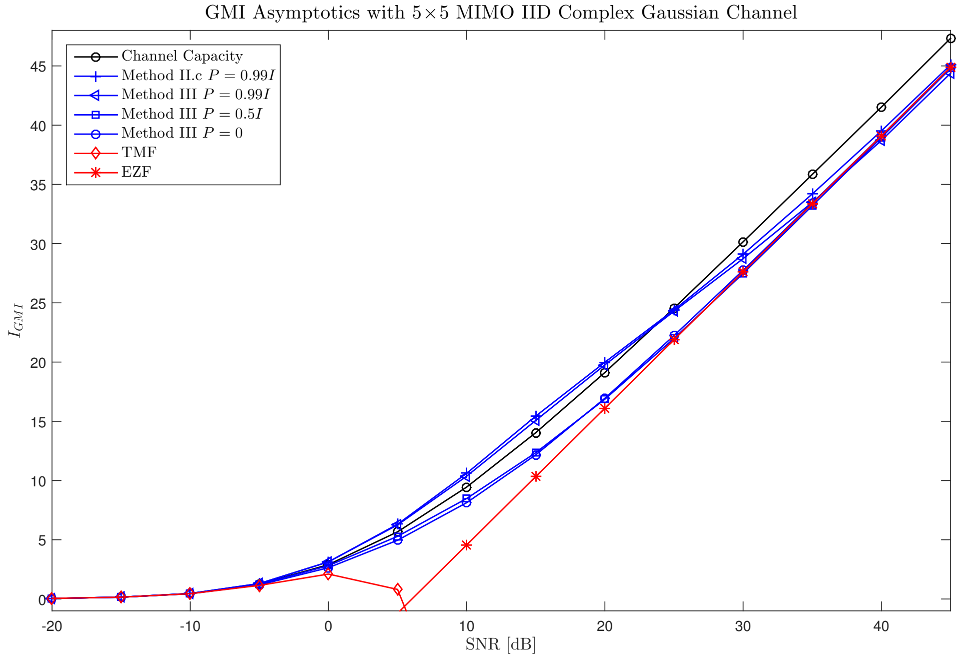

Fig. 8: GMI SNR asymptotic results of 5×5 MIMO IID complex Gaussian channels with $\nu = 1$.

II.c performs the best while Method I is quite close to it. Method II.b and Method III are inferior. Although it is not consistent with the GMI simulation results, where Method II.b has the highest GMI, it can be well explained: The GMI only measures the achievable rate under ideal conditions that optimal detection and decoding are utilized which is impractical of most practical systems. Further, the GMI is evaluated under the assumption that the pair $(\boldsymbol{x}, \hat{\boldsymbol{x}})$ is jointly Gaussian which is not the case in practice. Therefore the GMI can not fully reflect the final performance. However within the same receiver structure (Method I, Method II.b and Method II.c have different receiver structures since the CS parameters are not the same), when the feedback quality increases, the GMIs also increase for each individual method. Nevertheless, the BLER results in Figure 10 reveal two interesting facts: Firstly, through iterations with optimized CS parameters, the BLER performance improves dramatically and secondly, in Method II the interference cancelation matrix $\boldsymbol{R}$ shall be zero inside the band where $\boldsymbol{G}$ has non-zero elements.

Next in Figure 11 we plot the performance for the Proakis-B channel with 16QAM symbols and a (1064, 1920) turbo code. The conclusions are almost the same as with QPSK symbols, except that Method I performs slightly better than Method II.c with two and three iterations.

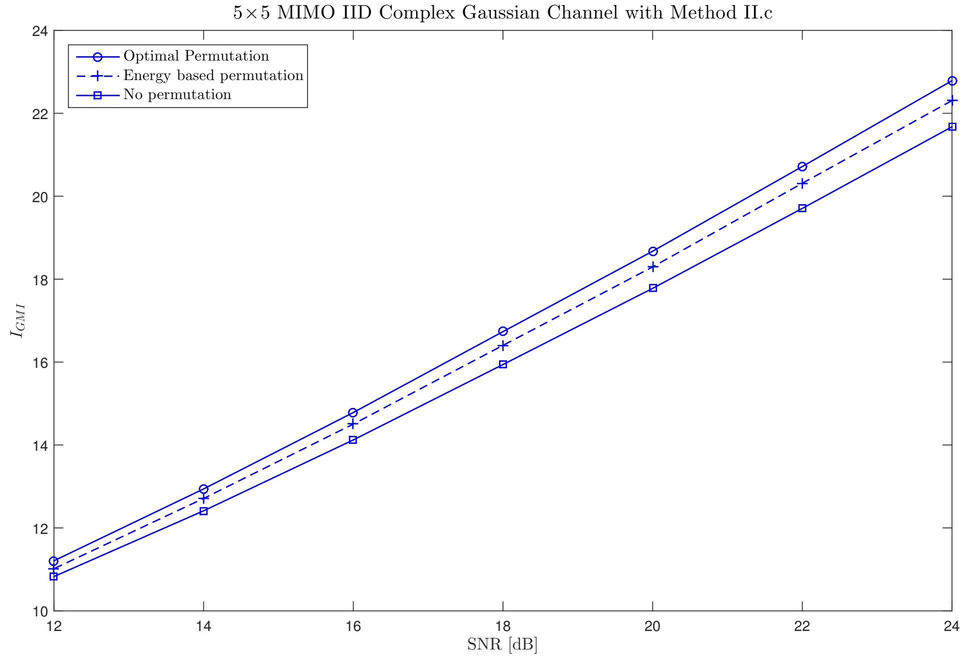

Fig. 9: Optimal permutation results for 5×5 MIMO IID complex Gaussian channels with Method II.c and $\nu = 1$.

In Figure 12 we plot the performance for the EPR4 [46] channel with QPSK symbols and a (1064, 1600) turbo code. The four taps are $\boldsymbol{h} = [\, 0.5 \; 0.5 \; -0.5 \; -0.5 \,]$. In this case, the LMMSE-PIC demodulator is quite efficient as with three iterations the performance gap compared with the MAP demodulator is around 1.5 dB at 10^{-1} BLER. We tested the CS demodulators both for $\nu = 1$ and $\nu = 2$. With $\nu = 1$, the performance gain of CS demodulators over LMMSE-PIC is around 1 dB at the first iteration and 0.4dB after three iterations at 10^{-1} BLER. With $\nu = 2$, the performance gain of the CS demodulators over LMMSE-PIC is more than 1 dB after three iterations and the gap compared with the MAP is less than 0.5 dB. For all CS demodulators the performance is quite close to each other, but Method II.c is slightly better than the others.

In Figure 13 we plot the performance for the EPR4 channel with 16QAM symbols and a (1064, 1920) turbo code. The performance of all CS demodulators is quite close to each other and outperform LMMSE-PIC with more than 1 dB. Moreover, Method I performs slightly better than the others after three iterations at high SNR.

In Figure 14 we plot the performance for the Proakis-C channel with QPSK symbols and a (1064, 1620) turbo code. The five taps are $\boldsymbol{h} = [\, 0.227 \; 0.46 \; 0.688 \; 0.46 \; 0.227 \,]$. We test the CS

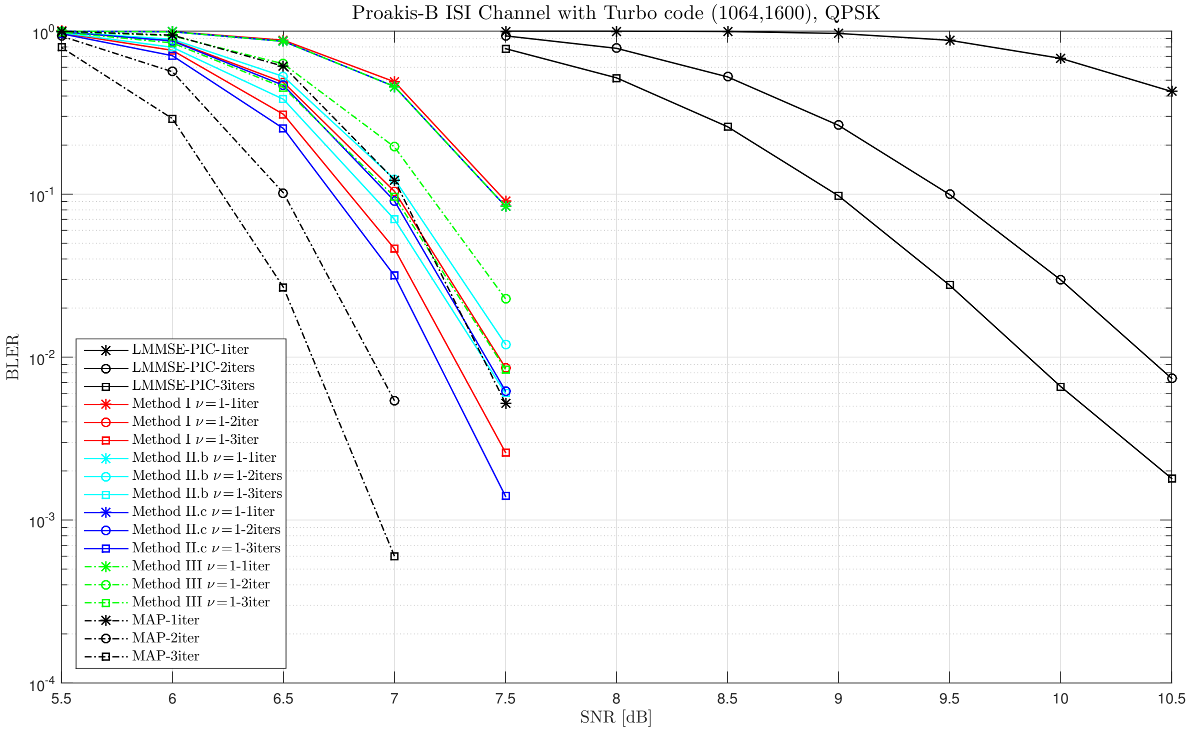

Fig. 10: Performance evaluation with Proakis-B channel with QPSK Symbols.

demodulators both for $\nu=1$ and $\nu=2$. The conclusions are similar to the other channels that have been tested with QPSK symbols. Method II.c performs better than Method I and Method II.b while Method III is slightly inferior. With $\nu=1$, the CS demodulators outperform LMMSE-PIC with more than 2 dB in all three iterations. With $\nu=2$, the gap compared with MAP is reduced to less than 1 dB while LMMSE-PIC has a gap to MAP that is up to 10 dB.

C. *MIMO Channels with Turbo Code*

Next we evaluate the BLER performance for MIMO channel. An interesting case would be when the receive antenna number is less than the transmit antenna number, i.e., $N<K$ in (1). In this case the LMMSE-PIC demodulator will fail [47] at the first iteration due to the lack of receive diversity. We tested 4×6 MIMO IID complex Gaussian channels with QPSK symbols and a (1064, 1800) turbo code. In Figure 15 we plot the performance comparison with 1, 2 and 3 global iterations. As can be seen, Method II.c performs better than Method I, Method II.a and Method II.b both for $\nu=1$ (right figure) and $\nu=3$ (left figure). In the first iteration, the performance of all methods are almost overlapped with each other and cannot be easily

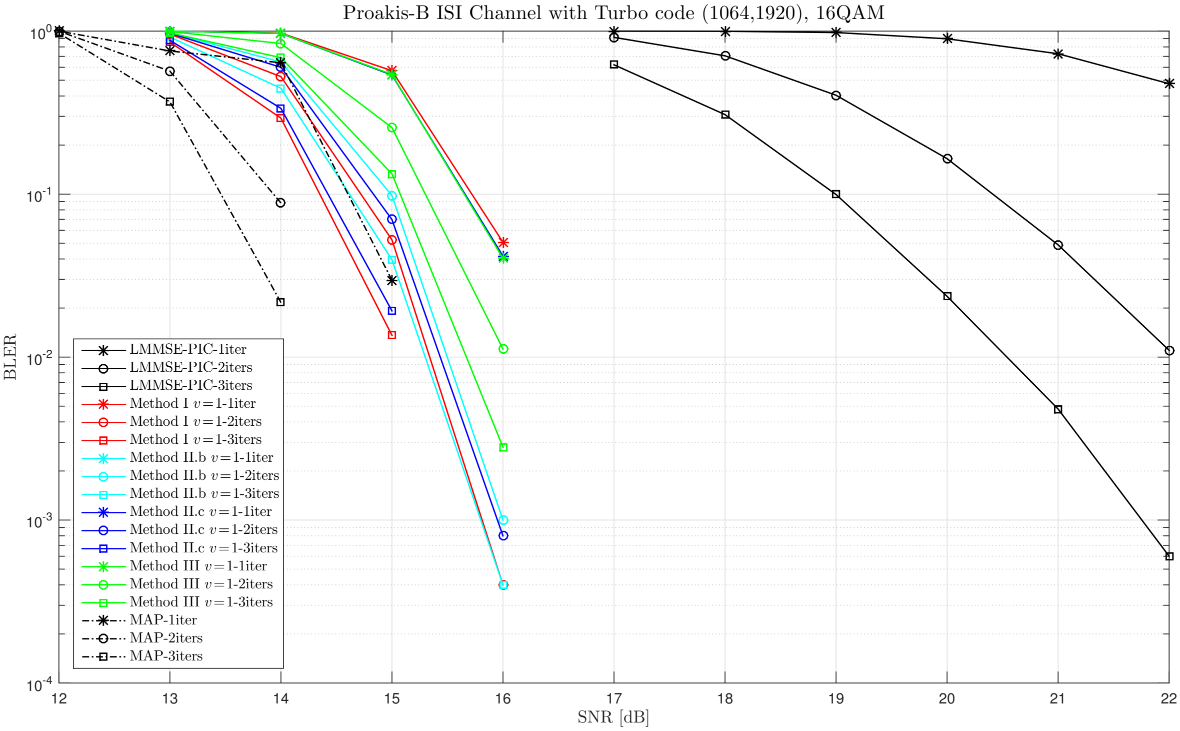

Fig. 11: Performance evaluation with Proakis-B channel with 16QAM Symbols.

distinguished from the plot.

The performance of the LMMSE-PIC and MAP demodulators are compared with Method II.c and Method III in Figure 16. LMMSE-PIC is inferior, especially at the first iteration where there is no soft information available. Method II.c with $\nu = 1$ improves the performance more than 1 dB over LMMSE-PIC with the same number of iterations. With $\nu = 3$, Method II.c is quite close to MAP, with less than 1 dB gap at 10^{-1} BLER. Again in both cases Method III performs quite close to Method II.c.

The impact of permuting the channel matrix is also simulated with Method II.c with $\nu = 1$. It can be seen that, in Figure 17 the energy based permutation outperforms the performance with no permutation, which is aligned with the GMI simulations that are presented in Figure 9.

Finally we remark that, for the sake of complexity savings, both for finite linear vector channels and ISI channels, the parameters of CS demodulators do not need to be updated through all iterations. Once the feedback information quality is good enough and the parameter $\boldsymbol{P}$ or α are close to ideal, the optimal CS parameters can be kept unchanged in remaining iterations.

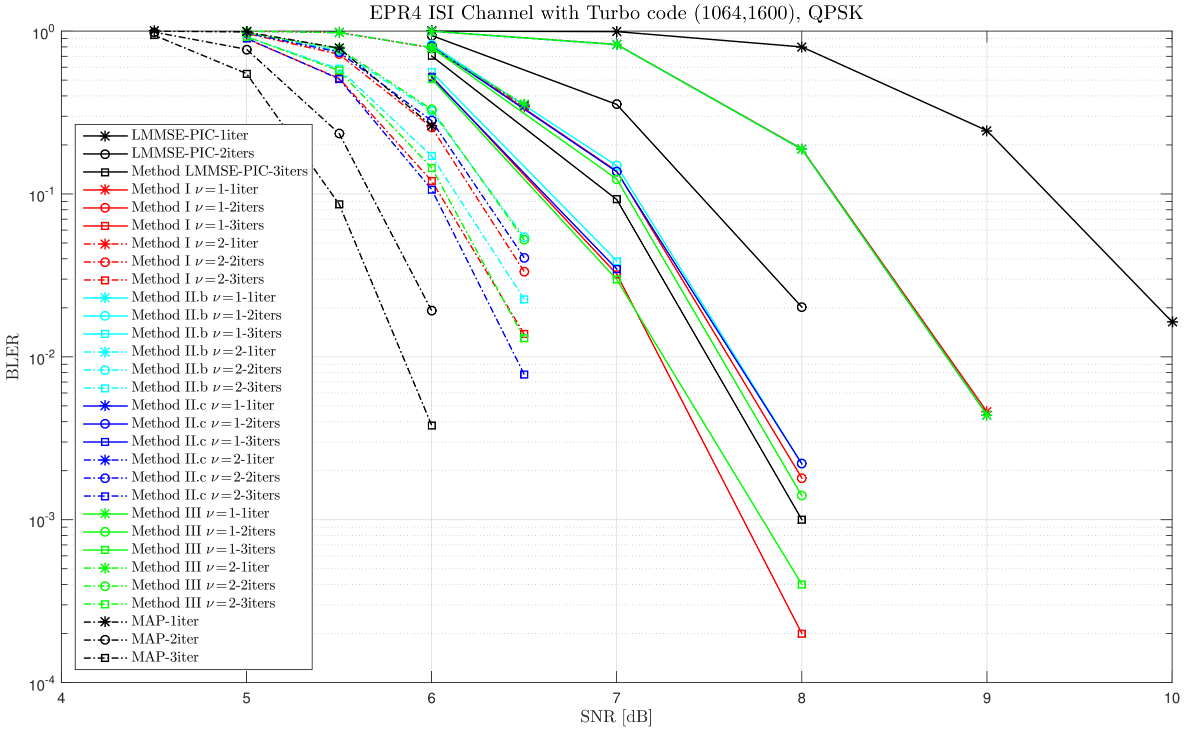

Fig. 12: Performance evaluation with EPR4 channel with QPSK Symbols.

VIII. Summary

In this paper we considered the design of CS demodulators for linear channels that use a trellis representation of the received signal in combination with interference cancelation of the signal part that is not appropriately modeled by the trellis. In order to reach a trellis representation, a linear filter is applied as front-end. It is an extension of the well studied CS demodulators to iterative receivers. We analyzed the properties of three different approaches for designing such optimal CS demodulators. In the used framework, there are three parameters that need to be optimized. Based on a generalized mutual information cost function, two of these are solved for in closed form, while the third needs to be numerically optimized except for the last method where we constructed it explicitly at the cost of a small performance loss. A simple gradient based optimization is used and turns out to perform well. Numerical results are provided to illustrate the behavior of the proposed CS demodulators. In general, Method II.c which is based on the Ungerboeck model outperforms Method I that is based on the Forney model. However Method I performs slightly better than Method II in some cases with 16QAM symbols. Method

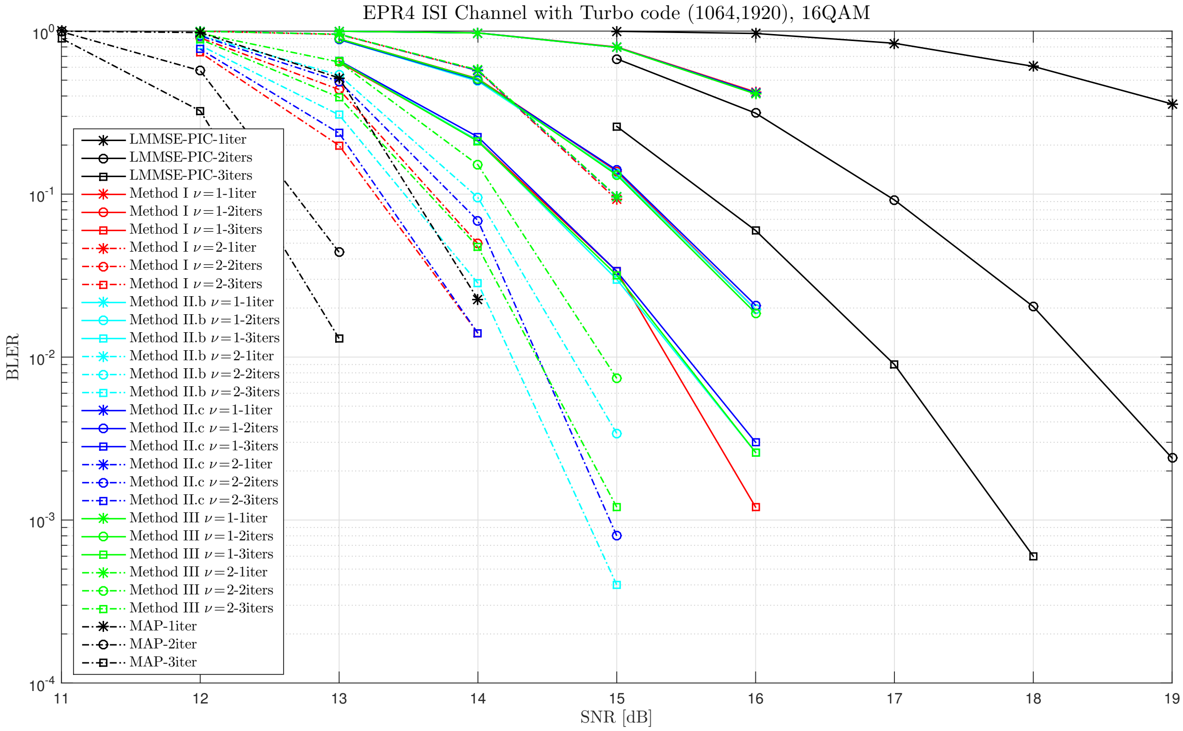

Fig. 13: Performance evaluation with EPR4 channel with 16QAM Symbols.

II has the advantage over Method I that the optimization procedure is concave. Furthermore, the suboptimal Method III performs close to Method I and Method II while it has all parameters in closed form. An interesting result is that the interference cancelation matrix should not cancel the effective channel perfectly outside the memory length. We also analyzed asymptotic properties and showed that Method III converges to Method II asymptotically when the noise density goes to zero or infinity.

Appendix A: Derivation of the GMI

By making the eigenvalue decomposition $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{H}}=\boldsymbol{G}$ and letting $\boldsymbol{s}=\boldsymbol{Q}^{\mathrm{H}}\boldsymbol{x}$. As $\boldsymbol{x}$ is assumed to be zero mean complex Gaussian random vector with covariance matrix $\boldsymbol{I}$, we can write $\tilde{p}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}})$ in (3) as

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}}) = \exp\big(2\mathrm{Re}\big\{\boldsymbol{s}^{\mathrm{H}}\boldsymbol{d}\big\}-\boldsymbol{s}^{\mathrm{H}}\boldsymbol{\Lambda}\boldsymbol{s}\big), \tag{78}$$

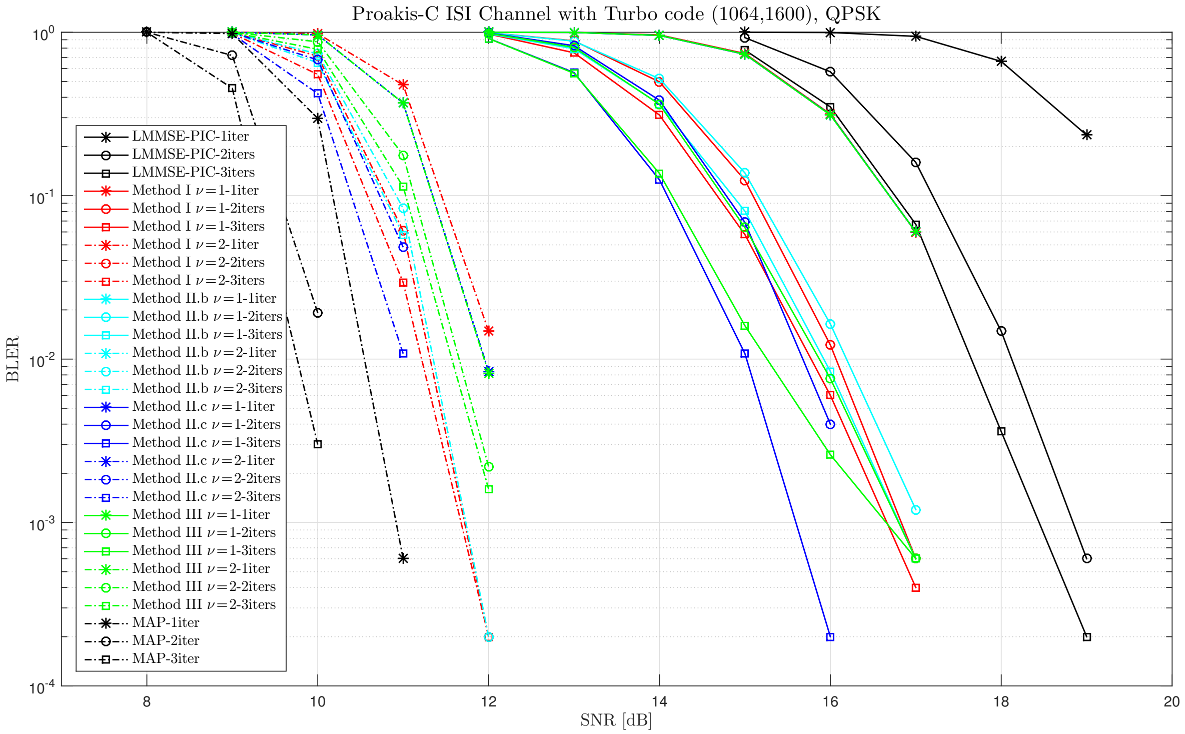

Fig. 14: Performance evaluation with Proakis-C channel with QPSK Symbols.

where $\boldsymbol{d}=\boldsymbol{Q}^{\rm H}(\boldsymbol{V}\boldsymbol{y}-\boldsymbol{R}\hat{\boldsymbol{x}})$. We can now evaluate

$$\begin{aligned}\tilde{p}(\boldsymbol{y}|\hat{\boldsymbol{x}}) &= \int \tilde{p}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \frac{1}{\pi^K}\int \exp\big(2\mathrm{Re}\{\boldsymbol{s}^{\rm H}\boldsymbol{d}\}-\boldsymbol{s}^{\rm H}\boldsymbol{\Lambda}\boldsymbol{s}\big)\exp\big(-\boldsymbol{s}^{\rm H}\boldsymbol{s}\big)\mathrm{d}\boldsymbol{s}\\ &= \prod_{k=1}^{N}\frac{1}{1+\lambda_k}\exp\left(\frac{|d_k|^2}{1+\lambda_k}\right).\end{aligned}$$

where λ_k is the kth diagonal element of $\boldsymbol{\Lambda}$ and d_k is the kth entry of $\boldsymbol{d}$. Taking the average over $\boldsymbol{y}$ and $\hat{\boldsymbol{x}}$ gives

$$-\mathbb{E}[\log(\tilde{p}(\boldsymbol{y}|\hat{\boldsymbol{x}}))] = \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)-\mathrm{Tr}\big(\boldsymbol{L}(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)$$

where the matrix $\boldsymbol{L}=\mathbb{E}[\boldsymbol{Q}\boldsymbol{d}\boldsymbol{d}^{\rm H}\boldsymbol{Q}^{\rm H}]$ is given by

$$\boldsymbol{L} = \boldsymbol{V}(N_0\boldsymbol{I}+\boldsymbol{H}\boldsymbol{H}^{\rm H})\boldsymbol{V}^{\rm H}-\boldsymbol{V}\boldsymbol{H}\boldsymbol{P}\boldsymbol{R}^{\rm H}-\boldsymbol{R}\boldsymbol{P}\boldsymbol{H}^{\rm H}\boldsymbol{V}^{\rm H}+\boldsymbol{R}\boldsymbol{P}\boldsymbol{R}^{\rm H}.$$

On the other hand, we have

$$-\mathbb{E}\left[\log(\tilde{p}(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{x}}))\right]=\mathrm{Tr}(\boldsymbol{G})-2\mathrm{Re}\big\{\mathrm{Tr}(\boldsymbol{V}\boldsymbol{H}-\boldsymbol{R}\boldsymbol{P})\big\}.$$

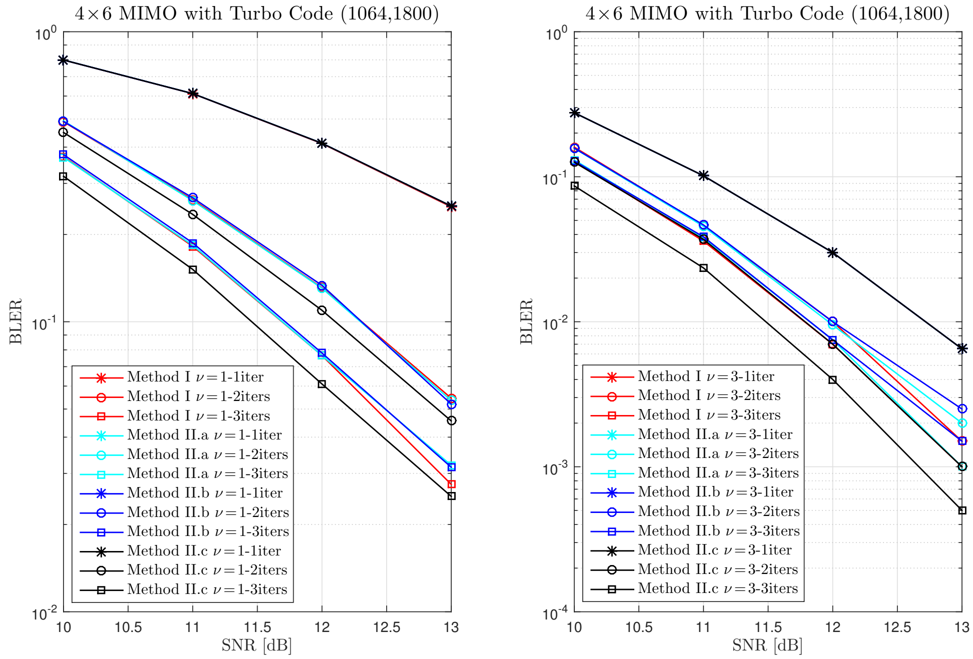

Fig. 15: Performance evaluation between Method I, Method II.a, Method II.b and Method II.c.

Combining the two expectations, the GMI reads,

$$\begin{aligned}I_{\text{GMI}}(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G}) &= \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)-\text{Tr}\big(\boldsymbol{L}(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)-\text{Tr}(\boldsymbol{G})+2\text{Re}\big\{\text{Tr}(\boldsymbol{V}\boldsymbol{H}-\boldsymbol{R}\boldsymbol{P})\big\}\\ &= \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)-\text{Tr}(\boldsymbol{G})+2\text{Re}\big\{\text{Tr}(\boldsymbol{V}\boldsymbol{H}-\boldsymbol{R}\boldsymbol{P})\big\}\\ &\quad -\text{Tr}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\big(\boldsymbol{V}[\boldsymbol{H}\boldsymbol{H}^{\text{H}}+N_0\boldsymbol{I}]\boldsymbol{V}^{\text{H}}-2\text{Re}\big\{\boldsymbol{V}\boldsymbol{H}\boldsymbol{P}\boldsymbol{R}^{\text{H}}\big\}+\boldsymbol{R}\boldsymbol{P}\boldsymbol{R}^{\text{H}}\big)\big).\end{aligned}$$

Appendix B: The Proof of Proposition 1

As the formula of GMI in (8) is quadratic in $\boldsymbol{W}$ and no constraints apply to $\boldsymbol{W}$, taking the gradient of $I_{\text{GMI}}(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F})$ with respect to $\boldsymbol{W}$ and setting it to zero, the optimal $\boldsymbol{W}$ is given in (11). Inserting $\boldsymbol{W}_{\text{opt}}$ into (8) gives, after some manipulations,

$$\begin{aligned}I_{\text{GMI}}(\boldsymbol{W}_{\text{opt}},\boldsymbol{T},\boldsymbol{F}) &= K+\log\big(\det(\boldsymbol{I}+\boldsymbol{F}^{\text{H}}\boldsymbol{F})\big)+\text{Tr}\big(\boldsymbol{T}^{\text{H}}\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\text{H}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\text{H}}\boldsymbol{T}\tilde{\boldsymbol{M}}\big)\\ &\quad +\text{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{F}^{\text{H}}\boldsymbol{F})\big)+2\text{Re}\big\{\text{Tr}\big(\boldsymbol{P}\boldsymbol{M}\boldsymbol{F}^{\text{H}}\boldsymbol{T}\big)\big\}.\end{aligned} \tag{79}$$

where $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$ are defined in (9) and (10).

If $\boldsymbol{P}=\boldsymbol{0}$, (79) equals

$$I_1(\boldsymbol{F}) = K+\log\big(\det(\boldsymbol{I}+\boldsymbol{F}^{\text{H}}\boldsymbol{F})\big)+\text{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{F}^{\text{H}}\boldsymbol{F})\big).$$

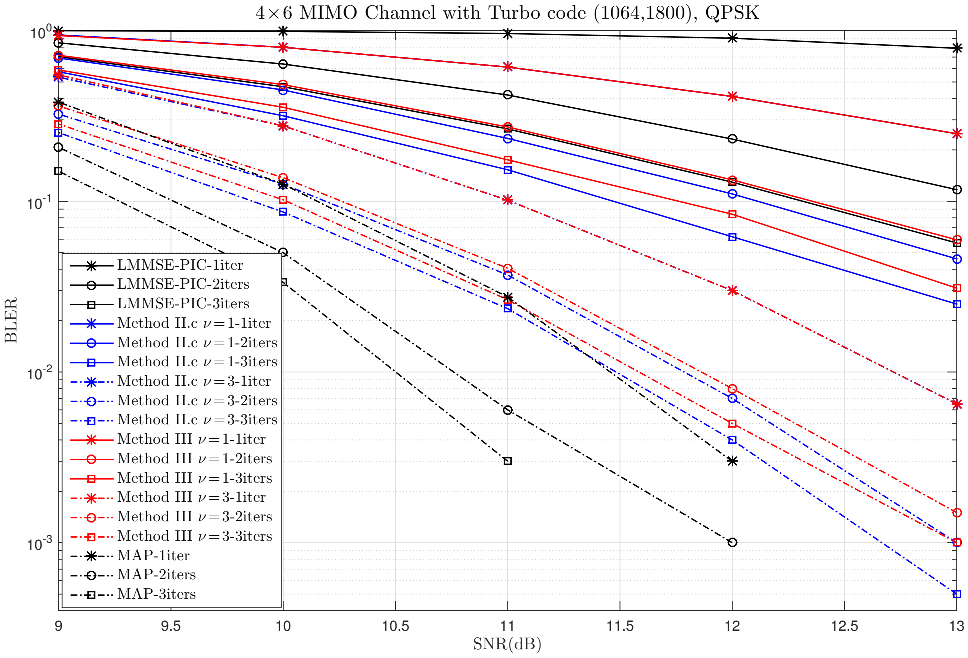

Fig. 16: Performance evaluation between the LMMSE-PIC, Method II.c and the MAP.

In this case, there is no soft information available and the matrix $\boldsymbol{T}$ is not included in the formula. When $\boldsymbol{P}\neq\boldsymbol{0}$, the terms of I_{GMI} in (79) related to $\boldsymbol{T}$ are

$$f(\boldsymbol{T})=\operatorname{Tr}\big(\boldsymbol{T}^{\text{H}}\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\text{H}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\text{H}}\boldsymbol{T}\tilde{\boldsymbol{M}}\big)+2\text{Re}\big\{\operatorname{Tr}\big(\boldsymbol{P}\boldsymbol{M}\boldsymbol{F}^{\text{H}}\boldsymbol{T}\big)\big\}.$$

Let $\boldsymbol{t}_k$ denote the kth column of $\boldsymbol{T}$, but all elements in rows $[k,\min(k+\nu,K-1)]$ removed, and define the column vector $\boldsymbol{t}=[\,\boldsymbol{t}_0^{\text{T}}\ \boldsymbol{t}_1^{\text{T}}\ \ldots\ \boldsymbol{t}_{K-1}^{\text{T}}]^{\text{T}}$, then by the definition of the indication matrix $\boldsymbol{\Omega}$, we have

$$\boldsymbol{t}=\boldsymbol{\Omega}\text{vec}(\boldsymbol{T}).$$

Similarly, let $\boldsymbol{z}_k$ denote the kth column of the matrix $\boldsymbol{F}\boldsymbol{M}\boldsymbol{P}$ but with all elements in rows $[k,\min(k+\nu,K-1)]$ are removed, and define the row vector $\boldsymbol{z}=[\,\boldsymbol{z}_0^{\text{T}}\ \boldsymbol{z}_1^{\text{T}}\ \ldots\ \boldsymbol{z}_{K-1}^{\text{T}}]^{\text{T}}$, then we have

$$\boldsymbol{z}=\Omega\text{vec}(\boldsymbol{F}\boldsymbol{M}\boldsymbol{P})=\boldsymbol{\Omega}\big((\boldsymbol{P}\boldsymbol{M}^{*})\otimes\boldsymbol{I}_K\big)\text{vec}(\boldsymbol{F}).$$

Finally, define a Hermitian matrix $\hat{\boldsymbol{B}}_1$ as

$$\hat{\boldsymbol{B}}_1=\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*}\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\text{H}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\text{H}}\big)\big)\boldsymbol{\Omega}^{\text{T}},$$

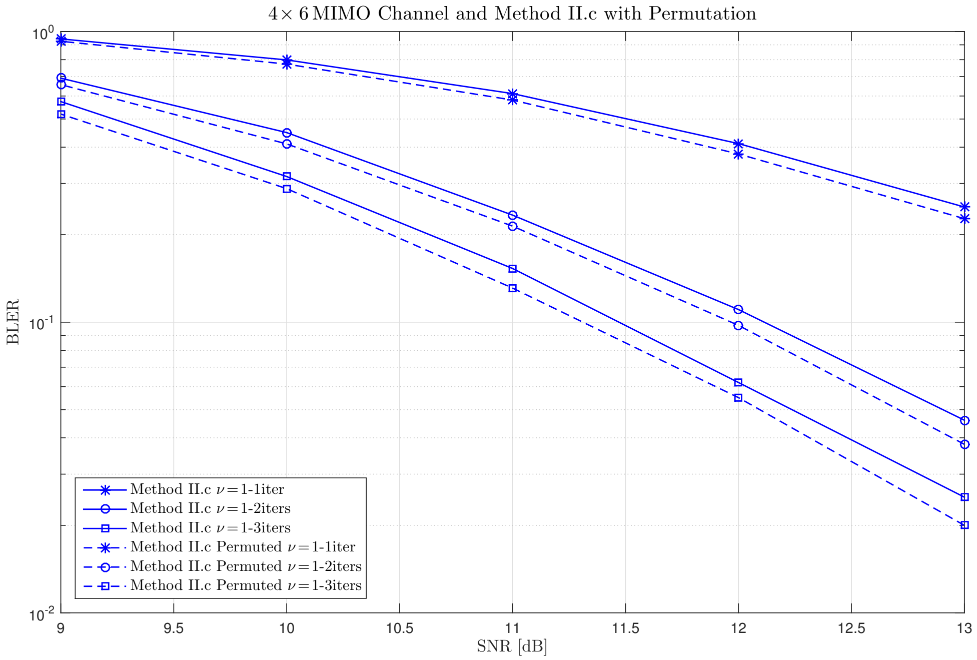

Fig. 17: Performance evaluation with channel permutations of Method II.c.

with that, we can rewrite $f(\boldsymbol{T})$ as

$$f(\boldsymbol{T}) = \boldsymbol{t}^{\mathrm{H}}\hat{\boldsymbol{B}}_1\boldsymbol{t}+2\mathrm{Re}\{\boldsymbol{z}^{\mathrm{H}}\boldsymbol{t}\}.$$

Taking the gradient with respect to $\boldsymbol{t}$ and setting it to zero yields,

$$\boldsymbol{t}_{\mathrm{opt}} = -\hat{\boldsymbol{B}}_1^{-1}\boldsymbol{z}. \tag{80}$$

Transferring $\boldsymbol{t}_{\mathrm{opt}}$ back into $\boldsymbol{T}_{\mathrm{opt}}$ given the optimal $\boldsymbol{T}$ in (12) and inserting this into $f(\boldsymbol{T})$ gives

$$f(\boldsymbol{T}_{\mathrm{opt}}) = -\boldsymbol{z}^{\mathrm{H}}\hat{\boldsymbol{B}}_1^{-1}\boldsymbol{z}.$$

Thus, with the optimal $\boldsymbol{W}$ and $\boldsymbol{T}$, when $\boldsymbol{P}\neq\boldsymbol{0}$ the GMI equals

$$\begin{aligned} I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T}_{\mathrm{opt}},\boldsymbol{F}) = {} & K+\log\big(\det(\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})\big) \\ & -\mathrm{vec}(\boldsymbol{F})^{\mathrm{H}}\boldsymbol{D}^{\mathrm{H}}\Big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*}\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{H}}\big)\big)\boldsymbol{\Omega}^{\mathrm{T}}\Big)^{-1}\boldsymbol{D}\mathrm{vec}(\boldsymbol{F}). \end{aligned}$$

where $\boldsymbol{D}=\boldsymbol{\Omega}\big((\boldsymbol{P}\boldsymbol{M}^{*})\otimes\boldsymbol{I}_K\big)$.

Appendix C: Non-Concavity Examples of Method I

We give examples to demonstrate the non-concavity of Method I for MIMO and ISI channels with assuming that $\boldsymbol{P}=\boldsymbol{I}$ and $\alpha=1$, respectively. The memory length $\nu=1$ and the noise density N_0 equals 1 in both cases. A 5×5 MIMO channel and the Proakis-C channel are used.

Example 4. *MIMO:*

$$\boldsymbol{H}=\begin{bmatrix} 2 & 0 & -3 & 5 & 4\\ -5 & 2 & -1 & 0 & 2\\ 2 & -4 & 3 & 3 & 3\\ -1 & -5 & -4 & 1 & 2\\ 0 & -2 & 0 & 5 & 5 \end{bmatrix},\boldsymbol{F}_1=\begin{bmatrix} 4.94 & 4.45 & 0 & 0 & 0\\ 0 & 0.21 & 3.85 & 0 & 0\\ 0 & 0 & 5.56 & 1.76 & 0\\ 0 & 0 & 0 & 0.61 & 7.10\\ 0 & 0 & 0 & 0 & 2.79 \end{bmatrix},\boldsymbol{F}_2=\begin{bmatrix} 2.03 & 6.17 & 0 & 0 & 0\\ 0 & 5.22 & 3.56 & 0 & 0\\ 0 & 0 & 7.43 & 0.73 & 0\\ 0 & 0 & 0 & 4.98 & 4.32\\ 0 & 0 & 0 & 0 & 10.11 \end{bmatrix}.$$

Example 5. *ISI:*

$$\boldsymbol{h}=\begin{bmatrix}0.227 & 0.460 & 0.688 & 0.460 & 0.227\end{bmatrix},\boldsymbol{f}_1=\begin{bmatrix}0.1606 & 0.9009\end{bmatrix},\boldsymbol{f}_2=\begin{bmatrix}0.2230 & 0.2035\end{bmatrix}.$$

The $I_{\text{GMI}}(\boldsymbol{W}_{\text{opt}},\boldsymbol{T}_{\text{opt}},\boldsymbol{F})$ given in (13) as a function $\boldsymbol{F}$ is plotted on the left in Figure 18, while the $\bar{I}(W_{\text{opt}}(\omega),T_{\text{opt}}(\omega),F(\omega))$ given in (51) as a function of $F(\omega)$ is plotted on the right. If $I_{\text{GMI}}(\boldsymbol{W}_{\text{opt}},\boldsymbol{T}_{\text{opt}},\boldsymbol{F})$ and $\bar{I}(W_{\text{opt}}(\omega),T_{\text{opt}}(\omega),F(\omega))$ are concave or convex, the blue curves lie above or below the black curves, which clearly does not hold in our examples.

Appendix D: Derivation of the Gradient in Method I with Finite Linear Vector Channel

In this section we derive the first order differential of the GMI given in (13) with respect to $\boldsymbol{F}$. In order to utilize the differential with respect to a matrix, we use the α-differential as defined in [48]. Assume a matrix $\boldsymbol{Y}_{N,K}$ with dimension $N\times K$ and a matrix $\boldsymbol{X}_{M,S}$ with dimension $M\times S$, define $d_{\boldsymbol{X}}\boldsymbol{Y}$ as the α-differential of $\boldsymbol{Y}$ with respect to $\boldsymbol{X}$. Furthermore, define y_ℓ and x_ℓ as $[\,y_1\ \ y_2\ \ \cdots\ \ y_{NK}\,]=\text{vec}(\boldsymbol{Y})^{\text{T}}$ and $[\,x_1\ \ x_2\ \ \cdots\ \ x_{MS}\,]=\text{vec}(\boldsymbol{X})^{\text{T}}$, the α-differential $d_{\boldsymbol{X}}\boldsymbol{Y}$ is defined as

$$d_{\boldsymbol{X}}\boldsymbol{Y}=\frac{\partial\text{vec}(\boldsymbol{Y})}{\partial\text{vec}(\boldsymbol{X})^{\text{T}}}=\begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_{MS}}\\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_{MS}}\\ \vdots & \vdots & \vdots & \vdots\\ \frac{\partial y_{NK}}{\partial x_1} & \frac{\partial y_{NK}}{\partial x_2} & \cdots & \frac{\partial y_{NK}}{\partial x_{MS}} \end{bmatrix}.$$

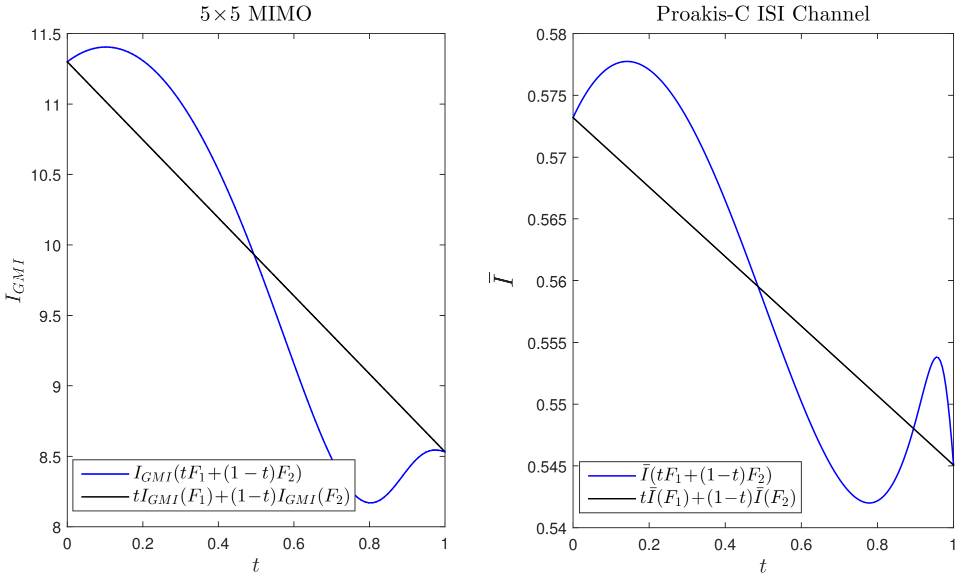

Fig. 18: Examples for the non-concaveness of Method I for 5×5 MIMO channel and Proakis-C ISI channel.

The reason for adopting the α-differential is because it keeps the chain rule and the product rule. At first we introduce an $NK\times NK$ permutation matrix $\boldsymbol{Z}_{N,K}$, which satisfies the condition

$$\mathrm{vec}(\boldsymbol{Y}^{\mathrm{T}}) = \boldsymbol{Z}_{N,K}\mathrm{vec}(\boldsymbol{Y}).$$

It is easy to verify that $\boldsymbol{Z}_{N,K}^{-1}=\boldsymbol{Z}_{K,N}$, and when $N=1$ or $K=1$, $\boldsymbol{Y}$ is a vector and $\mathrm{vec}(\boldsymbol{Y}^{\mathrm{T}})=\mathrm{vec}(\boldsymbol{Y})$, hence $\boldsymbol{Z}_{N,1}=\boldsymbol{I}_N$ and $\boldsymbol{Z}_{1,K}=\boldsymbol{I}_K$. Furthermore, by definition we have

$$d_{\boldsymbol{F}}(\boldsymbol{F})=d_{\boldsymbol{F}}(\mathrm{vec}(\boldsymbol{F}))=\boldsymbol{I}$$
$$d_{\boldsymbol{F}}(\boldsymbol{F}^{\mathbf{H}})=d_{\boldsymbol{F}}\left(\mathrm{vec}(\boldsymbol{F}^{\mathrm{H}})\right)=\mathbf{0}.$$

We start by reviewing a few properties [48], [49] of α-differential that will be used later. Assume that both matrix $\boldsymbol{X}$ and $\boldsymbol{Y}$ are functions of $\boldsymbol{F}$ and the dimensions are specified by the subscripts associated to them, the below equations hold:

$$\begin{aligned}
d_{\boldsymbol{F}}(\boldsymbol{X}_{K,K}^{-1}) &= -(\boldsymbol{X}_{K,K}^{-\mathrm{T}}\otimes\boldsymbol{X}_{K,K}^{-1})d_{\boldsymbol{F}}\boldsymbol{X}_{K,K}\\
d_{\boldsymbol{F}}(\boldsymbol{Y}_{N,K}\boldsymbol{X}_{K,S}) &= (\boldsymbol{X}_{K,S}^{\mathrm{T}}\otimes\boldsymbol{I}_N)d_{\boldsymbol{F}}\boldsymbol{Y}_{N,K}+(\boldsymbol{I}_S\otimes\boldsymbol{Y}_{N,K})d_{\boldsymbol{F}}\boldsymbol{X}_{K,S}\\
d_{\boldsymbol{F}}(\log(\det(\boldsymbol{X}_{K,K}))) &= \mathrm{vec}(\boldsymbol{X}_{K,K}^{-\mathrm{T}})^{\mathrm{T}}d_{\boldsymbol{F}}\boldsymbol{X}_{K,K}\\
d_{\boldsymbol{F}}(\boldsymbol{Y}_{N,K}\otimes\boldsymbol{X}_{M,S}) &= (\boldsymbol{I}_K\otimes\boldsymbol{Z}_{S,N}\otimes\boldsymbol{I}_M)(\boldsymbol{I}_{NK}\otimes\mathrm{vec}(\boldsymbol{X})d_{\boldsymbol{F}}\boldsymbol{Y}_{N,K}\\
&\quad+(\boldsymbol{I}_K\otimes\boldsymbol{Z}_{S,N}\otimes\boldsymbol{I}_M)(\mathrm{vec}(\boldsymbol{Y})\otimes\boldsymbol{I}_{MS})d_{\boldsymbol{F}}\boldsymbol{X}_{M,S}.
\end{aligned}$$

The α-differential of $I_1(\boldsymbol{F})$ with respect to $\boldsymbol{F}$ is

$$\begin{aligned} d_{\boldsymbol{F}}(I_1) &= \mathrm{vec}((\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})^{-\mathrm{T}})^{\mathrm{T}}(\boldsymbol{I}_K\otimes\boldsymbol{F}^{\mathrm{H}})+\mathrm{vec}(\boldsymbol{F}^{*}\boldsymbol{M}^{\mathrm{T}})^{\mathrm{T}} \\ &= \mathrm{vec}(\boldsymbol{F}\boldsymbol{M}+\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})^{-\mathrm{H}})^{\mathrm{H}}. \end{aligned} \tag{81}$$

Define a $K\times K$ matrix $\boldsymbol{B}=\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{H}}$ and an $S\times S$ matrix $\mathbf{\Pi}=\big(\Omega(\tilde{\boldsymbol{M}}^{*}\otimes\boldsymbol{B})\Omega^{\mathrm{T}}\big)^{-1}$, the α-differential of $\delta_1(\boldsymbol{F})$ with respect to $\boldsymbol{F}$ is

$$d_{\boldsymbol{F}}(\delta_1) = -\mathrm{vec}(\boldsymbol{F})^{\mathrm{H}}\boldsymbol{D}^{\mathrm{H}}\left((\mathrm{vec}(\boldsymbol{F})^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}})\otimes\boldsymbol{I}_S\right)d_{\boldsymbol{F}}(\mathbf{\Pi})-\mathrm{vec}(\boldsymbol{F})^{\mathrm{H}}\boldsymbol{D}^{\mathrm{H}}\mathbf{\Pi}\boldsymbol{D}, \tag{82}$$

where

$$\begin{aligned} d_{\boldsymbol{F}}(\mathbf{\Pi}) &= -(\mathbf{\Pi}^{\mathrm{T}}\otimes\mathbf{\Pi})d_{\boldsymbol{F}}(\boldsymbol{\Omega}(\tilde{\boldsymbol{M}}^{*}\otimes\boldsymbol{B})\boldsymbol{\Omega}^{\mathrm{T}}) \\ &= -(\mathbf{\Pi}^{\mathrm{T}}\otimes\mathbf{\Pi})(\boldsymbol{\Omega}\otimes\boldsymbol{\Omega})d_{\boldsymbol{F}}(\tilde{\boldsymbol{M}}^{*}\otimes\boldsymbol{B}) \\ &= -\big((\mathbf{\Pi}^{\mathrm{T}}\boldsymbol{\Omega})\otimes(\mathbf{\Pi}\boldsymbol{\Omega})\big)(\boldsymbol{I}_K\otimes\boldsymbol{Z}_{K,K}\otimes\boldsymbol{I}_K)(\mathrm{vec}(\tilde{\boldsymbol{M}}^{\boldsymbol{*}})\otimes\boldsymbol{I}_{K^2})d_{\boldsymbol{F}}\boldsymbol{B} \end{aligned} \tag{83}$$

and

$$\begin{aligned} d_{\boldsymbol{F}}(\boldsymbol{B}) &= d_{\boldsymbol{F}}\big(\boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})^{-1}\big) \\ &= \big((\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})^{-\mathrm{T}}\big)\otimes\big((\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})^{-1}\big)d_{\boldsymbol{F}}(\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}) \\ &= \big((\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})^{-\mathrm{T}}\big)\otimes\big((\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})^{-1}\big)(\boldsymbol{F}^{*}\otimes\boldsymbol{I}_K) \\ &= \big(\boldsymbol{F}^{*}(\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})^{-\mathrm{T}}\big)\otimes(\boldsymbol{I}-\boldsymbol{B}). \end{aligned} \tag{84}$$

Define a $K\times K$ matrix $\tilde{\boldsymbol{F}}=(\boldsymbol{I}+\boldsymbol{F}^{\mathrm{H}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{H}}$ and a $K^4\times K^2$ matrix

$$\boldsymbol{\Psi} = (\boldsymbol{I}_K\otimes\boldsymbol{Z}_{K,K}\otimes\boldsymbol{I}_K)\big(\mathrm{vec}(\tilde{\boldsymbol{M}}^{\boldsymbol{*}})\otimes\boldsymbol{I}_{K^2}\big), \tag{85}$$

by combing (81)-(85), we finally have, when $\boldsymbol{P}\neq\boldsymbol{0}$,

$$\begin{aligned} d_{\boldsymbol{F}}\big(I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T}_{\mathrm{opt}},\boldsymbol{F})\big) &= d_{\boldsymbol{F}}(I_1)+d_{\boldsymbol{F}}(\delta_1) \\ &= \mathrm{vec}\big(\boldsymbol{F}\boldsymbol{M}+\tilde{\boldsymbol{F}}^{\mathrm{H}}\big)^{\mathrm{H}}-\mathrm{vec}(\boldsymbol{F})^{\mathrm{H}}\boldsymbol{D}^{\mathrm{H}}\mathbf{\Pi}\boldsymbol{D} \\ &\quad +\mathrm{vec}(\boldsymbol{F})^{\mathrm{H}}\boldsymbol{D}^{\mathrm{H}}\big((\boldsymbol{\Omega}^{\mathrm{T}}\mathbf{\Pi}\boldsymbol{D}\mathrm{vec}(\boldsymbol{F}))^{\mathrm{T}}\otimes\big(\mathbf{\Pi}\boldsymbol{\Omega}\big)\big)\boldsymbol{\Psi}\big(\tilde{\boldsymbol{F}}^{\mathrm{T}}\otimes(\boldsymbol{I}-\boldsymbol{B})\big). \end{aligned}$$

Appendix E: The Proof of Proposition 3

As the formula of GMI in (6) is quadratic in $\boldsymbol{V}$ and no constraints apply to $\boldsymbol{V}$, taking the gradient of $I_{\mathrm{GMI}}(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G})$ with respect to $\boldsymbol{V}$ and setting it to zero, yields the optimal $\boldsymbol{V}$ given in (21). Inserting $\boldsymbol{V}_{\mathrm{opt}}$ into (6) gives, after some manipulations

$$\begin{aligned} I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R},\boldsymbol{G}) &= K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+2\mathrm{Re}\big\{\mathrm{Tr}(\boldsymbol{P}\boldsymbol{M}\boldsymbol{R})\big\} \\ &\quad +\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big)+\mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\boldsymbol{R}\tilde{\boldsymbol{M}}\boldsymbol{R}^{\mathrm{H}}\big) \end{aligned} \tag{86}$$

where $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$ are defined in (9) and (10).

If $\boldsymbol{P}=\boldsymbol{0}$, (86) equals

$$I_2(\boldsymbol{G}) = K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big).$$

When $\boldsymbol{P}\neq\boldsymbol{0}$, the terms of $I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R},\boldsymbol{G})$ in (86) related to $\boldsymbol{R}$ are

$$g(\boldsymbol{R})= 2\mathrm{Re}\big\{\mathrm{Tr}(\boldsymbol{P}\boldsymbol{M}\boldsymbol{R})\big\}+\mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\boldsymbol{R}\tilde{\boldsymbol{M}}\boldsymbol{R}^{\mathrm{H}}\big).$$

Let $\boldsymbol{r}_k$ denote the kth column of $\boldsymbol{R}$, but where all elements in rows $[\max(0,k-\nu_{\mathrm{R}}),\min(k+\nu_{\mathrm{R}},K-1)]$ are removed, and define the column vector $\boldsymbol{r}=[\,\boldsymbol{r}_0^{\mathrm{T}}\;\boldsymbol{r}_1^{\mathrm{T}}\;\ldots\;\boldsymbol{r}_{K-1}^{\mathrm{T}}]^{\mathrm{T}}$, then we have

$$\boldsymbol{r} = \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{R}).$$

Moreover, let $\boldsymbol{d}_k$ denote the kth column of the matrix $\boldsymbol{M}\boldsymbol{P}$ but with all elements in rows $[\max(0,k-\nu_{\mathrm{R}}),\min(k+\nu_{\mathrm{R}},K-1)]$ are removed and define the vector $\boldsymbol{d}=[\boldsymbol{d}_0^{\mathrm{T}}\;\boldsymbol{d}_1^{\mathrm{T}}\;\ldots\;\boldsymbol{d}_{K-1}^{\mathrm{T}}]^{\mathrm{T}}$. From the definition of $\boldsymbol{d}$, we have

$$\boldsymbol{d} = \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{M}\boldsymbol{P}).$$

Defining a Hermitian matrix $\hat{\boldsymbol{B}}_2$ as

$$\hat{\boldsymbol{B}}_2 = \boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*}\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}},$$

we can write $f(\boldsymbol{R})$ as

$$g(\boldsymbol{R}) = \boldsymbol{r}^{\mathrm{H}}\hat{\boldsymbol{B}}_2\boldsymbol{r} + 2\,\mathrm{Re}\{\boldsymbol{d}^{\mathrm{H}}\boldsymbol{r}\}.$$

Therefore the optimal $\boldsymbol{r}$ is

$$\boldsymbol{r}_{\mathrm{opt}} = -\hat{\boldsymbol{B}}_2^{-1}\boldsymbol{d}. \tag{87}$$

Transferring $\boldsymbol{r}_{\text{opt}}$ back into $\boldsymbol{R}_{\text{opt}}$ gives the optimal $\boldsymbol{R}$ in (22) and inserting this into $g(\boldsymbol{R})$ gives

$$g(\boldsymbol{R}_{\text{opt}}) = -\boldsymbol{d}^{\text{H}} \hat{\boldsymbol{B}}_2^{-1} \boldsymbol{d}.$$

Thus, with the optimal $\boldsymbol{V}$ and $\boldsymbol{R}$, when $\boldsymbol{P} \neq \boldsymbol{0}$ the GMI equals

$$\begin{aligned} I_{\text{GMI}}(\boldsymbol{V}_{\text{opt}}, \boldsymbol{R}_{\text{opt}}, \boldsymbol{G}) = K + \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) + \operatorname{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big) \\ -\boldsymbol{d}^{\text{H}}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\text{T}}\big)^{-1}\boldsymbol{d}. \end{aligned}$$

Appendix F: Derivation of the Gradient in Method II with Finite Linear Vector Channel

Now we calculate the α-differential of $I_{\text{GMI}}(\boldsymbol{V}_{\text{opt}}, \boldsymbol{R}_{\text{opt}}, \boldsymbol{G})$ given in (23) with respect to $\boldsymbol{G}$ when $\boldsymbol{P} \neq \boldsymbol{0}$. Taking the α-differential of $I_2(\boldsymbol{G})$ with respect to $\boldsymbol{G}$ yields,

$$d_{\boldsymbol{G}}(I_2) = \operatorname{vec}((\boldsymbol{I}+\boldsymbol{G})^{-1} + \boldsymbol{M})^{\text{H}}. \tag{88}$$

Define an $S \times S$ Hermitian matrix $\boldsymbol{\Phi} = \big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\text{T}}\big)^{-1}$ and taking the α-differential of $\delta_2(\boldsymbol{G})$ with respect to $\boldsymbol{G}$ yields,

$$\begin{aligned} d_{\boldsymbol{G}}(\delta_2) &= -(\boldsymbol{d}^{\text{T}} \otimes \boldsymbol{d}^{\text{H}}) d_{\boldsymbol{G}}(\boldsymbol{\Phi}) \\ &= (\boldsymbol{d}^{\text{T}} \otimes \boldsymbol{d}^{\text{H}})(\boldsymbol{\Phi}^{\text{T}} \otimes \boldsymbol{\Phi}) d_{\boldsymbol{G}}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\text{T}}\big) \\ &= \big((\boldsymbol{d}^{\text{T}}\boldsymbol{\Phi}^{\text{T}}) \otimes (\boldsymbol{d}^{\text{H}}\boldsymbol{\Phi})\big)(\boldsymbol{\Omega} \otimes \boldsymbol{\Omega}) d_{\boldsymbol{G}}\big(\tilde{\boldsymbol{M}}^{*} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big) \\ &= \big((\boldsymbol{d}^{\text{T}}\boldsymbol{\Phi}^{\text{T}}\boldsymbol{\Omega}) \otimes (\boldsymbol{d}^{\text{H}}\boldsymbol{\Phi}\boldsymbol{\Omega})\big)\boldsymbol{\Psi} d_{\boldsymbol{G}}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\big) \\ &= -\big((\boldsymbol{d}^{\text{T}}\boldsymbol{\Phi}^{\text{T}}\boldsymbol{\Omega}) \otimes (\boldsymbol{d}^{\text{H}}\boldsymbol{\Phi}\boldsymbol{\Omega})\big)\boldsymbol{\Psi}\big((\boldsymbol{I}+\boldsymbol{G})^{-\text{T}} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big) \end{aligned} \tag{89}$$

where $\boldsymbol{\Psi}$ is defined in (85). Combining (88) and (89), we can obtain

$$\begin{aligned} d_{\boldsymbol{G}}\big(I_{\text{GMI}}(\boldsymbol{V}_{\text{opt}}, \boldsymbol{R}_{\text{opt}}, \boldsymbol{G})\big) &= d_{\boldsymbol{G}}(I_2) + d_{\boldsymbol{G}}(\delta_2) \\ &= \operatorname{vec}\big((\boldsymbol{I}+\boldsymbol{G})^{-1} + \boldsymbol{M}\big)^{\text{H}} \\ &\quad -\big((\boldsymbol{d}^{\text{T}}\boldsymbol{\Phi}^{\text{T}}\boldsymbol{\Omega}) \otimes (\boldsymbol{d}^{\text{H}}\boldsymbol{\Phi}\boldsymbol{\Omega})\big)\boldsymbol{\Psi}\big((\boldsymbol{I}+\boldsymbol{G})^{-\text{T}} \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big). \end{aligned}$$

Appendix G: The Concavity Proof of Method II with Finite Linear Vector Channels

When $\boldsymbol{P}=\boldsymbol{0}$, as $\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)$ is concave [50] and $\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big)$ is linear in $\boldsymbol{G}$, the function $I_2(\boldsymbol{G})$ in (24) is concave with respect to $\boldsymbol{G}$ whenever $\boldsymbol{I}+\boldsymbol{G}$ is positive definite.

The concavity when $\boldsymbol{P}\neq\boldsymbol{0}$ can be deduced from the composition theorem in [50, Chpater 3.6]. For a positive definite matrix $\boldsymbol{X}$, $\boldsymbol{d}^{\mathrm{H}}\boldsymbol{X}^{-1}\boldsymbol{d}$ is convex and non-increasing (with respect to the generalized inequality for positive definite Hermitian matrices, see [50], [51]) for any column vector $\boldsymbol{d}$. Furthermore, since $\boldsymbol{I}+\boldsymbol{G}$ is positive definite, $(\boldsymbol{I}+\boldsymbol{G})^{-1}$ is convex. As $\tilde{\boldsymbol{M}}\prec\boldsymbol{0}$ $\boldsymbol{X}=\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*}\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}$ is concave in $\boldsymbol{G}$.

By the composition theorem, $\boldsymbol{d}^{\mathrm{H}}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^{*}\otimes[\boldsymbol{I}+\boldsymbol{G}]^{-1}\big)\boldsymbol{\Omega}^{\mathrm{T}}\big)^{-1}\boldsymbol{d}$ is convex, and $\delta_2(\boldsymbol{G})$ is then concave. Therefore the function $I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R}_{\mathrm{opt}},\boldsymbol{G})$ in (23) is concave with respect to $\boldsymbol{G}$ whenever $\boldsymbol{I}+\boldsymbol{G}$ is positive definite.

Appendix H: The Proof of Proposition 5

The Fourier series associated to the Toeplitz matrix $\boldsymbol{W}$ is

$$W(\omega)=\sum_{k=-\infty}^{\infty}w_k\exp(jk\omega)\,,$$

and the differential of $\bar{I}(W(\omega),T(\omega),F(\omega))$ in (44) with respect to w_k (where ω is fixed) is

$$\begin{aligned}\frac{\partial\bar{I}}{\partial w_k}=&-\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|F(\omega)|^2\big(N_0+|H(\omega)|^2\big)W^*(\omega)}{1+|F(\omega)|^2}\exp(jk\omega)\,\mathrm{d}\omega\\&+\frac{1}{\pi}\int_{-\pi}^{\pi}\Big(F^*(\omega)H(\omega)+\frac{\alpha|F(\omega)|^2H(\omega)T^*(\omega)}{1+|F(\omega)|^2}\Big)\exp\big(jk\omega\big)\mathrm{d}\omega.\end{aligned}\tag{90}$$

Since (90) should equal zero for all k, the optimal $W(\omega)$ is given in (49). Putting $W_{\mathrm{opt}}(\omega)$ back into (44) yields,

$$\begin{aligned}\bar{I}(W_{\mathrm{opt}}(\omega),T(\omega),F(\omega))=1+\frac{\alpha}{\pi}\int_{-\pi}^{\pi}\mathrm{Re}\big\{F^*(\omega)T(\omega)M(\omega)\big\}\mathrm{d}\omega+\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+|F(\omega)|^2\big)\\+\frac{\tilde{M}(\omega)|T(\omega)F(\omega)|^2}{1+|F(\omega)|^2}+M(\omega)\big(1+|F(\omega)|^2\big)\Big)\mathrm{d}\omega.\end{aligned}\tag{91}$$

where $M(\omega)$ and $\tilde{M}(\omega)$ are defined in (45) and (46).

When $\alpha=0$, the GMI in (91) equals (52) and when $0<\alpha\leq1$, the terms related to $T(\omega)$ in (91) are

$$f(T(\omega))=\frac{\alpha}{\pi}\int_{-\pi}^{\pi}\mathrm{Re}\big\{F^*(\omega)T(\omega)M(\omega)\big\}\mathrm{d}\omega+\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|T(\omega)F(\omega)|^2}{1+|F(\omega)|^2}\mathrm{d}\omega.\tag{92}$$

As the elements of the main diagonal and the first ν lower diagonals of matrix $\boldsymbol{T}$ are constrained to zero, we define the vector $\tilde{\boldsymbol{t}}$ that specifies the Toeplitz matrix $\boldsymbol{T}$ as

$$\tilde{\boldsymbol{t}}=[\, t_{-N_\mathrm{T}} \ \ldots \ t_{-1} \, t_{\nu+1} \ \ldots \ t_{N_\mathrm{T}} \,],$$

and with $\phi(\omega)$ defined in (47), the Fourier series $T(\omega)$ with a finite tap length N_T is

$$T(\omega)=\sum_{-N_\mathrm{T}\leq k\leq N_\mathrm{T}, k\notin[0,\nu]} t_k \exp\big(jk\omega\big)=\tilde{\boldsymbol{t}}\phi(\omega). \tag{93}$$

Furthermore, with $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ defined in (48), (92) can be rewritten as

$$f(T(\omega))=\tilde{\boldsymbol{t}}\boldsymbol{\varepsilon}_2\tilde{\boldsymbol{t}}^\mathrm{H}+2\mathrm{Re}\big\{\tilde{\boldsymbol{t}}\boldsymbol{\varepsilon}_1\big\}.$$

Therefore the optimal $\tilde{\boldsymbol{t}}$ is

$$\tilde{\boldsymbol{t}}_\mathrm{opt}=-\boldsymbol{\varepsilon}_1^\mathrm{H}\boldsymbol{\varepsilon}_2^{-1}. \tag{94}$$

Putting $\tilde{\boldsymbol{t}}_\mathrm{opt}$ back into (91)-(93), the optimal $T(\omega)$ is given in (50) and $\bar{I}(W(\omega),T(\omega),F(\omega))$ for the optimal $W(\omega)$ and $T(\omega)$ is given in (51).

Appendix I: The Proof of Proposition 7

The Fourier series associated to the Toeplitz matrix $\boldsymbol{V}$ is

$$V(\omega)=\sum_{k=-\infty}^{\infty} v_k \exp(jk\omega)\,,$$

and the differential of $\bar{I}(V(\omega),R(\omega),G(\omega))$ in (56) with respect to v_k (where ω is fixed) is

$$\begin{aligned}\frac{\partial\bar{I}}{\partial v_k}=&-\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\big(N_0+|H(\omega)|^2\big)V^*(\omega)}{1+G(\omega)}\exp(jk\omega)\,\mathrm{d}\omega\\&+\frac{1}{\pi}\int_{-\pi}^{\pi}\Big(H(\omega)+\frac{\alpha H(\omega)R^*(\omega)}{1+G(\omega)}\Big)\exp\big(jk\omega\big)\mathrm{d}\omega.\end{aligned} \tag{95}$$

Since (95) shall equal zero for all k, the optimal $V(\omega)$ is given in (59). Putting $V_\mathrm{opt}(\omega)$ in (59) back into (56) yields,

$$\begin{aligned}\bar{I}(V_\mathrm{opt}(\omega),R(\omega),G(\omega))=&\ 1+\frac{\alpha}{\pi}\int_{-\pi}^{\pi}\mathrm{Re}\big\{M(\omega)R(\omega)\big\}\mathrm{d}\omega+\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+G(\omega)\big)\\&+\frac{\tilde{M}(\omega)|R(\omega)|^2}{1+G(\omega)}+M(\omega)\big(1+G(\omega)\big)\Big)\mathrm{d}\omega,\end{aligned} \tag{96}$$

where $M(\omega)$ and $\tilde{M}(\omega)$ are defined in (45) and (46).

When $\alpha = 0$, the GMI in (96) equals (62) and when $0 < \alpha \leq 1$, the terms of $\bar{I}(V_{\rm opt}(\omega), R(\omega), G(\omega))$ related to $R(\omega)$ in (96) are

$$g(R(\omega)) = \frac{\alpha}{\pi}\int_{-\pi}^{\pi} \mathrm{Re}\big\{M(\omega)R(\omega)\big\}\mathrm{d}\omega + \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\tilde{M}(\omega)|R(\omega)|^2}{1+G(\omega)}\mathrm{d}\omega. \tag{97}$$

Define the vector $\tilde{\boldsymbol{r}}$ that specifies the Toeplitz matrix $\boldsymbol{R}$ as

$$\tilde{\boldsymbol{r}} = [\, r_{-N_{\rm R}} \, \ldots \, r_{-\nu_{\rm R}-1} \; r_{\nu_{\rm R}+1} \, \ldots \, r_{N_{\rm R}}],$$

and with $\psi(\omega)$ defined in (57), the Fourier series $R(\omega)$ with a finite tap length $N_{\rm R}$ is

$$R(\omega) = \sum_{-N_{\rm R} \leq k \leq N_{\rm R}, k \notin [-\nu_{\rm R}, \nu_{\rm R}]} r_k \exp\big(jk\omega\big) = \tilde{\boldsymbol{r}}\psi(\omega) \tag{98}$$

where $2\nu_{\rm R}+1$ is the band size that $\boldsymbol{R}$ is constrained to zero. With $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$ defined in (58), (97) can be written as

$$g(R(\omega)) = \tilde{\boldsymbol{r}}\boldsymbol{\zeta}_2\tilde{\boldsymbol{r}}^{\rm H} + 2\mathrm{Re}\left\{\tilde{\boldsymbol{r}}\boldsymbol{\zeta}_1\right\}.$$

Therefore the optimal $\tilde{\boldsymbol{r}}$ is

$$\tilde{\boldsymbol{r}}_{\rm opt} = -\boldsymbol{\zeta}_1^{\rm H}\boldsymbol{\zeta}_2^{-1}. \tag{99}$$

This shows that $\tilde{\boldsymbol{r}}_{\rm opt}$ has Hermitian symmetry as $G(\omega)$, $M(\omega)$ and $\tilde{M}(\omega)$ are all real valued, thus $R_{\rm opt}(\omega)$ is real. Putting $\tilde{\boldsymbol{r}}_{\rm opt}$ back into (96)-(98), the optimal $R(\omega)$ is given in (60) and $\bar{I}(V(\omega), R(\omega), G(\omega))$ for the optimal $V(\omega)$ and $R(\omega)$ is given in (61).

Appendix J: The Concavity Proof of Method II with ISI Channels

In order to prove that $\bar{I}(V_{\rm opt}(\omega), R_{\rm opt}(\omega), G(\omega))$ in (61) is concave with respect to $G(\omega)$, it is sufficient to prove that $\boldsymbol{\zeta}_1^{\rm H}\boldsymbol{\zeta}_2^{-1}\boldsymbol{\zeta}_1$ is convex with respect to $G(\omega)$. For a positive definite matrix $\boldsymbol{\zeta}_2$, $\boldsymbol{\zeta}_1^{\rm H}\boldsymbol{\zeta}_2^{-1}\boldsymbol{\zeta}_1$ is convex and non-increasing (with respect to a generalized inequality for positive definite Hermitian matrices) in $G(\omega)$ for any vector $\boldsymbol{\zeta}_1$ and with arbitrary finite tap length $N_{\rm R}$. As matrix $\tilde{\boldsymbol{M}}$ is negative definite, $\boldsymbol{\zeta}_2$ in (58) is concave with respect to $G(\omega)$ under the constraint that $\boldsymbol{I}+\boldsymbol{G}$ is positive definite. Hence $\boldsymbol{\zeta}_1^{\rm H}\boldsymbol{\zeta}_2^{-1}\boldsymbol{\zeta}_1$ is convex in $G(\omega)$ by the composition theorem [50, Chapter 3.6].

APPENDIX K: THE PROOF OF LEMMA 5

With the GMI in Method III given in (38), from Theorem 2 the optimal $\boldsymbol{G}$ satisfies,

$$[(\boldsymbol{I}+\boldsymbol{G}_{\text{opt}})^{-1}]_{\nu}= -[\hat{\boldsymbol{M}}]_{\nu}.$$

Notice that, when $\boldsymbol{P}=\boldsymbol{0}$, Method III and Method II are equivalent since $\hat{\boldsymbol{M}}=\boldsymbol{M}$. Hence in order to prove Lemma 4, it is sufficient to prove that $[\hat{\boldsymbol{M}}]_{\nu}$ converges to $[\boldsymbol{M}]_{\nu}$ as $N_0\to 0$ and ∞ in Method III.

When $\boldsymbol{P}\prec\boldsymbol{I}$, $\boldsymbol{C}_k$ is positive definite as in (33) and as $N_0\to 0$,

$$\begin{aligned}\boldsymbol{H}^{\mathrm{H}}(\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}^{\mathrm{H}}+N_0\boldsymbol{I})^{-1}\boldsymbol{H} &= \boldsymbol{C}_k^{-1}(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}+N_0\boldsymbol{C}_k^{-1})^{-1}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}\\ &= \boldsymbol{C}_k^{-1}\big(\boldsymbol{I}-N_0\boldsymbol{C}_k^{-1}(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1}\big)+\mathcal{O}(N_0^2).\end{aligned}$$

Therefore with $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{C}}$ defined in (32)-(36),

$$\begin{aligned}\hat{\boldsymbol{W}}\boldsymbol{H} &= \boldsymbol{I}-N_0(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1}+\mathcal{O}(N_0^2),\\ \hat{\boldsymbol{C}} &= [\hat{\boldsymbol{W}}\boldsymbol{H}]_{\backslash\nu} = -N_0[(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1}]_{\backslash\nu}+\mathcal{O}(N_0^2).\end{aligned} \tag{100}$$

With (100) and $\hat{\boldsymbol{M}}$ defined in (39), it can be verified that,

$$\lim_{N_0\to 0}[\hat{\boldsymbol{M}}/N_0]_{\nu}=-[(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})^{-1}]_{\nu}.$$

On the other hand, when $N_0\to\infty$, from (32)-(39),

$$\begin{aligned}N_0\hat{\boldsymbol{W}} &= \boldsymbol{H}^{\mathrm{H}}(\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}^{\mathrm{H}}/N_0+\boldsymbol{I})^{-1} = \boldsymbol{H}^{\mathrm{H}}+\mathcal{O}(1/N_0),\\ N_0\hat{\boldsymbol{C}} &= [\hat{\boldsymbol{W}}\boldsymbol{H}]_{\backslash\nu} = [\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}]_{\backslash\nu}+\mathcal{O}(1/N_0),\end{aligned} \tag{101}$$

With (101) and $\hat{\boldsymbol{M}}$ defined in (39), it can be verified that,

$$\lim_{N_0\to\infty}[N_0(\boldsymbol{I}+\hat{\boldsymbol{M}})]_{\nu} = [\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}]_{\nu}.$$

Therefore, from (72) $[\hat{\boldsymbol{M}}]_{\nu}$ converges to $[\boldsymbol{M}]_{\nu}$ as $N_0\to 0$ and ∞, which completes the proof.

REFERENCES

[1] G. D. Forney Jr., “Maximum likelihood sequence estimation of digital sequences in the presence of intersymbol interference,” *IEEE Trans. on Inform. Theory*, vol. 18, no. 3, pp. 363–378, May 1972.

[2] A. J. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Trans. on Inform. Theory*, vol. IT-13, no. 2, pp. 260-269, Apr. 1967.

[3] D. D. Falconer, and F. R. Magee, “Adaptive channel memory truncation for maximum likelihood sequence estimation,” *The Bell System Technical Journal*, vol..51, no. 9, pp. 1541-1562, Nov. 1973.

[4] S. A. Fredricsson, “Joint optimization of transmitter and receiver filter in digital PAM systems with a Viterbi detector,” *IEEE Trans. on Inform. Theory*, vol. IT-22, no. 2, pp. 200-210, Mar. 1976.

[5] C. T. Beare, “The choice of the desired impulse response in combined linear-Viterbi algorithm equalizers,” *IEEE Trans. on Communications*, vol. 26, pp. 1301-1307, 1978.

[6] N. Sundström, O. Edfors, P. Ödling, H. Eriksson, T. Koski, and P. O. Börjesson, “Combined linear-Viterbi equalizers - a comparative study and a minimax design,” *In Proc.* IEEE Vehicular Technology Conference (VTC), pp. 1263-1267 vol. 2, Stockholm, Sweden, Jun. 1994.

[7] N. Al-Dhahir, and J. M. Cioffi, “Efficiently computed reduced-parameter input-aided MMSE equalizers for ML detection: A unified approach,” *IEEE Trans. on Inform. Theory*, vol. 42, pp. 903-915, Apr. 1996.

[8] M. A. Lagunas, A. I. Perez-Neia, and J. Vidal, “Joint beamforming and Viterbi equalizer in wireless communications,” *In Proc.* Thirty-First Asilomar Conference on Signals, Systems & Computers, pp. 915-919, vol. 1, Pacific Grove, (CA) , Nov. 1997.

[9] S. A. Aldosari, S. A. Alshebeili, and A. M. Al-Sanie, “A new MSE approach for combined linear-Viterbi equalizers,” *In Proc.* IEEE Vehicular Technology Conference (VTC), pp. 1707-1711, vol. 3, Tokyo, Japan, May 2000.

[10] R. Venkataramani and S. Sankaranarayanan, “Optimal channel shortening equalization for MIMO ISI channels,” *In Proc.* IEEE Global Telecommunications (GLOBECOM), New Orleans, (LO), Dec. 2008.

[11] A. Shaheem, *Iterative detection for wireless communications*, Ph.D. thesis, School of Electrical, Electronic and Computer Engineering, University of Western Australia, 2008.

[12] U. L. Dang, W. H. Gerstacker, and D. T. M. Slock, “Maximum SINR prefiltering for reduced state trellis based equalization,” *In Proc.* IEEE International Conference on Communications (ICC), Kyoto, Japan, Jun. 2011.

[13] R. Venkataramani and M. F. Erden, “A posteriori equivalence: A new perspective for design of optimal channel shortening equalizers,” *arXiv:0710.3802v1*.

[14] I. Abou-Faycal and A. Lapidoth, “On the capacity of reduced complexity receivers for intersymbol interference channels”, *In Proc.* Conference on Information Sciences and Systems (CISS), Princeton University, pp. WA4 32 -37, Mar. 2000.

[15] N. Merhav, G. Kaplan, A. Lapidoth and S. Shamai, “On information rates for mismatched decoders,” *IEEE Trans. on Inform. Theory*, Nov. 1994.

[16] A. Ganti, A. Lapidoth, and I. E. Telatar, “Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit,” *IEEE Trans. on Inform. Theory*, vol. 46, pp. 2315-2328, Nov. 2000.

[17] M. R. McKay, I. B. Collings, and A. M. Tulino, “Achievable sum rate of MIMO MMSE receivers: A general analytic framework,” *IEEE Trans. on Inform. Theory*, vol. 56, no. 1, Jan. 2010.

[18] F. Rusek and A. Prlja, “Optimal channel shortening of MIMO and ISI channels,” *IEEE Trans. on Wireless Commun.,* vol. 11, no. 2, pp. 810–818, Feb. 2012.

[19] H. Weingarten, Y. Steinberg, and S. Shamai, “Gaussian codes and weighted nearest neighbor decoding in fading multiple-antenna channels,” *IEEE Trans. on Inform. Theory*, vol. 50, no. 8, Aug. 2004.

[20] F. Rusek, N. Al-Dhahir, and A. Gomaa, “A rate-maximizing channel-shortening detector with soft feedback side information,” *In Proc.* IEEE Global Telecommunications (GLOBECOM), Anaheim, (CA), Dec., 2012.

[21] A. Duel-Hallen and C. Heegard, “Delayed decision-feedback sequence estimation,” *IEEE Trans. on Commun.,* vol. 37, no. 5, pp. 428-436, May 1989.

[22] J. Hagenauer, “Source-controlled channel decoding,” *IEEE Trans. on Commun.,* vol. 43, no. 9, pp. 2449-2457, Sep. 1995.

[23] S. Ten Brink, “Convergence behavior of iteratively decoded parallel concatenated codes,” *IEEE Trans. Commun.,* vol. 49, no. 10, pp. 17271737, Oct. 2001.

[24] S. M. Kay, “Fundamentals of statistical signal processing, volume I: estimation theory,” Prentice Hall, Apr. 1993.

[25] B. M. Hochwald and S. Ten Brink, “Achieving near-capacity on a multiple-antenna channel,” *IEEE Trans. Commun.,* vol. 51, no. 3, pp. 389399, 2003.

[26] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, “Optimal decoding of linear codes for minimizing symbol error rate,” *IEEE Trans. on Inform. Theory*, vol. IT-20(2), pp. 284-287, Mar. 1974.

[27] C. Studer, S. Fateh, and D. Seethaler, “ASIC Implementation of soft-input soft-output MIMO detection using parallel interference cancellation,” *IEEE Journal of Solid-State Circuits*, vol. 46, no. 7, pp. 1754-1765, Jul. 2011.

[28] M. Witzke, S. Bro, F. Schreckenbach, and J. Hagenauer, “Iterative detection of MIMO signals with linear detectors,” *In Proc.* Asilomar Conf. on Signals, Systems and Computers (ACSSC), Monterey, (CA), pp. 289–293, Nov. 2002.

[29] J. Zhang, H. Nguyen, and G. Mandyam, “LMMSE-based iterative and turbo equalization methods for CDMA downlink channels,” *In Proc.* IEEE 6th Workshop Signal Process. Advances Wireless Commun., pp. 231-235, Jun. 2005.

[30] F. Rusek, D. Fertonani, “Bounds on the information rate of intersymbol interference channels based on mismatched receivers,” *IEEE Trans. on Inform. Theory*, vol. 58, No. 3, pp. 1470-1482, 2012.

[31] G. Caire and S. Shamai, “On the achievable throughput of a multiantenna Gaussian broadcast channel,” *IEEE Trans. on Inform. Theory*, vol. 49, no. 7, Jul. 2003.

[32] G. L.Turin, “An introduction to digital matched filters,” *Proc. of the IEEE*, vol. 64, pp. 1092-1112, Jul. 1972.

[33] G. Ungerboeck, “Adaptive maximum likelihood receiver for carrier-modulated data-transmission systems,” *IEEE Trans. Commun.,* vol. 22, pp. 624-636, May 1974.

[34] G. H. Golub and C. F. Van Loan, “Matrix computations,” 3rd ed. Baltimore, MD: Johns Hopkins, p. 51, 1996.

[35] A. Kavčić and J. M. F. Moura, “Matrix with banded inverses: algorithms and factorization of Gauss-Markov processes,” *IEEE Trans. on Inform. Theory*, vol. 46, no. 4, pp. 1495-1509, Jul. 2000.

[36] G. Colavolpe and A. Barbieri, “On MAP symbol detection for ISI channels using the Ungerboeck observation model,” IEEE Communications Letters, vol. 9, no. 8, pp. 720-722, Aug. 2005.

[37] F. Rusek, G. Colavolpe, and, C. Sundberg, “40 years with the Ungerboeck model: a Look at its potentialities [Lecture Notes],” Signal Processing Magazine, vol. 32, pp. 156-161, May 2015.

[38] F. Rusek, M. Loncar and A. Prlja, “A comparison of Ungerboeck and Forney models for reduced complexity lSI equalization,” Proc. GLOBECOM, Washington D.C., pp. 1431-1436, Dec. 2007.

[39] O. Edfors, M. Sandell, J. J. Van de Beek, S. K. Wilson and P. O. Borjesson, “OFDM channel estimation by singular value decomposition,” *IEEE Trans. on Commun.,* vol. 46, no. 7, pp. 931-939, 1998.

[40] W. Hirt, *Capacity and information rates of discrete-time channels with memory*, Ph.D thesis, no. ETH 8671, Inst. Signal and Information Processing, Swiss Federal Inst. Technol., Zurich, 1988.

[41] U. Grenander anf G. Szegö, *Toeplitz forms and their applications*, University of Calif. Press Berkeley and Los Angeles, 1958.

[42] R. M. Gray, “Toeplitz and circulant matrices: A review,” Foundations and trends in communications and information theory, vol. 2, No. 3, pp 155-239, 2006.

[43] J. G. Proakis and M. Salehi, *Digital communications*, McGraw-Hill international edition, fifth edition, 2008.

[44] M. Tüchler, A. Singer and R. Kötter, “Minimum mean squared error (MMSE) equalization using priors,” *IEEE Trans. on Signal Processing*, vol. 50, pp. 673-683, 2000.

[45] 3GPP TS 36.212: *Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding*, Release 12, V12.4.0, Mar. 2015.

[46] P. Kabal and S. Pasupathy, “Partial-response signaling,” *IEEE Trans. Commun.,* vol. COM-23, pp. 921934, Sep. 1975.

[47] F. Rusek and O. Edfors, “An information theoretic characterization of channel shortening receivers,” 47th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, (CA), pp. 2108-2112, Nov. 2013.

[48] J. R. Magnus, “On the concept of matrix derivative,” *Journal of Multivariate Analysis*, vol. 101, no. 9, pp. 2200-2206, Oct. 2001.

[49] P. L. Fackler, “Notes on Matrix Calculus,” *Available from http://www4.ncsu.edu/pfackler/ MatCalc.pdf*, North Carolina State University, Sep. 2005.

[50] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[51] C. Davis, “Notions generalizing convexity for functions defined on spaces of matrices,” *Proc. Symp. Pure Math.,* vol. 7, pp. 187-201, 1963.